




ARTICLE



<https://doi.org/10.1057/s41599-023-01680-4>

OPEN

# A decomposition of book structure through ousiometric fluctuations in cumulative word-time

Mikaela Irene Fudolig<sup>1</sup><sup>✉</sup>, Thayer Alshaabi<sup>1,2</sup>, Kathryn Cramer<sup>1</sup>, Christopher M. Danforth<sup>1,3</sup> & Peter Sheridan Dodds<sup>1,4</sup>

While quantitative methods have been used to examine changes in word usage in books, studies have focused on overall trends, such as the shapes of narratives, which are independent of book length. We instead look at how words change over the course of a book as a function of the number of words, rather than the fraction of the book, completed at any given point; we define this measure as “cumulative word-time”. Using ousiometrics, a reinterpretation of the valence–arousal–dominance framework of meaning obtained from semantic differentials, we convert text into time series of power and danger scores, with time corresponding to cumulative word-time. Each time series is then decomposed using empirical mode decomposition into a sum of constituent oscillatory modes and a non-oscillatory trend. By comparing the decomposition of the original power and danger time series with those derived from shuffled text, we find that shorter books exhibit only a general trend, while longer books have fluctuations in addition to the general trend. These fluctuations typically have a period of a few thousand words regardless of the book length or library classification code but vary depending on the content and structure of the book. Our findings suggest that, in the ousiometric sense, longer books are not expanded versions of shorter books, but rather are more similar in structure to a concatenation of shorter texts. Further, they are consistent with editorial practices that require longer texts to be broken down into sections, such as chapters. Our method also provides a data-driven denoising approach that works for texts of various lengths, in contrast to the more traditional approach of using large window sizes that may inadvertently smooth out relevant information, especially for shorter texts. Altogether, these results open up avenues for future work in computational literary analysis, particularly the possibility of measuring a basic unit of narrative.

<sup>1</sup>Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Core, University of Vermont, Burlington, VT, USA. <sup>2</sup>Advanced Bioimaging Center, UC Berkeley, Berkeley, CA, USA. <sup>3</sup>Department of Mathematics & Statistics, University of Vermont, Burlington, VT, USA. <sup>4</sup>Department of Computer Science, University of Vermont, Burlington, VT, USA. ✉email: [mikaela.fudolig@uvm.edu](mailto:mikaela.fudolig@uvm.edu)

## Introduction

The computational study of word usage in the text has mainly been confined to aggregates, particularly the frequency of words in large corpora and contextual co-occurrence (Corral et al., 2015; Dodds et al., 2020; Reimers and Gurevych, 2019; Ryland Williams et al., 2015; Vaswani et al., 2017). Only recently have there been quantitative studies on how word usage changes over the course of a single text, and these have focused on the structure of narratives. Studies on narratives have traditionally been made from a qualitative perspective, requiring human inputs to interpret text (Brown and Tu, 2020; Freytag, 1900; Genette, 1983; Phelan and Rabinowitz, 2012; Ricoeur, 1980; Vonnegut, 1999). While the world's literary experts are far from being supplanted by computational analyses—and arguably never will be—their attentions cannot be scaled to study very large corpora, for which quantitative techniques have been developed.

Gao et al. (2016) showed that sentiment in novels has long-range correlations across the length of the book, indicating that sentiment exhibits a structure that relates to the flow of the novel. Inspired by Vonnegut (1999, 2010, 2005), Reagan et al. (2016) examined how smoothed, “distantly measured” (Moretti, 2013) happiness scores (Dodds et al., 2011) vary across the length of a book. Reagan et al. found that these happiness time series across fiction books reveal six major emotional arcs, identified as Rags-to-riches (rise), tragedy (fall), Icarus (rise–fall), Vonnegut's man-in-a-hole (fall–rise), Cinderella (rise–fall–rise), and Oedipus (fall–rise–fall). More recently, Boyd et al. (2020) showed that the different parts of Freytag's dramatic arc (Freytag, 1900) can be differentiated by the dominant word usage. Articles and prepositions are heavily used in the exposition stage, where narrators establish the setting. Once the reader has an understanding of the context of the story, the plot progression can proceed with more pronouns and function words. As the story reaches a climax, more cognitive process words are used as the characters and narrator work through the conflict (Boyd et al., 2020). While nonfiction works such as TED talks, US Supreme Court decisions, and *The New York Times* articles follow similar patterns in plot progression as works of fiction, they differ in patterns of staging and cognitive processes.

Schmidt (2015) put forward the idea of “plot arcs”, where a text is seen as a path through topics that represent a multidimensional space. Similarly, Toubia et al. (2021) studied narratives as paths in high-dimensional word embedding space. Dividing a text into segments represented by corresponding average word embedding vectors, the narrative is then described as the path in word embedding space obtained by moving along consecutive segments. They identify three properties of these paths, namely the speed, volume, and circuitousness in word embedding space, and examine how well these properties can predict the financial success of movies and TV shows as well as the number of citations in academic papers.

Though these studies look at changes in word usage across text, they focus on the general shape of narrative progression from beginning to end. Narrative is often characterized as a time series, with the general arc represented by the rise and fall of certain markers (such as happiness scores, Reagan et al., 2016) throughout the plot in terms of the fraction of the book covered. This normalization by the text length allows for a direct comparison of texts of different lengths, such as a short story and a novel, and the corresponding shapes of their respective arcs. However, this approach does not allow us to compare sections of a text with other sections of comparable length, for example, a chapter of a novel vs. a short story of similar length.

To do this, not just for narratives but for any type of text, we have to use *cumulative word-time*, or simply *word-time*, which we

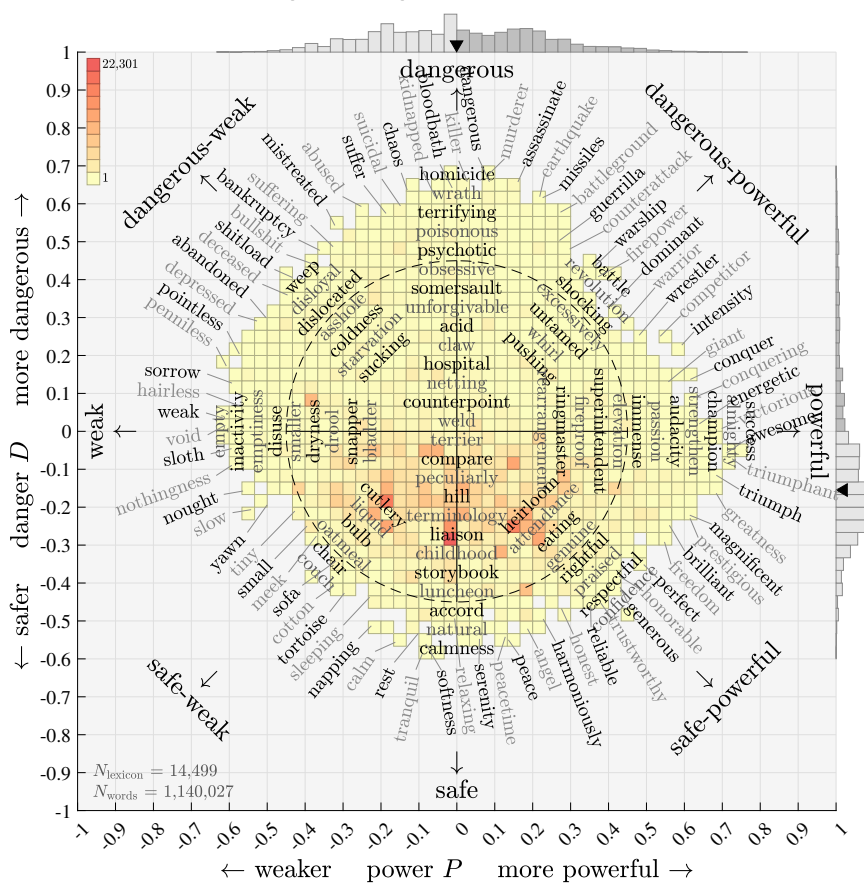
define as the number, rather than the fraction, of words covered in a text at a given point. Though it is numerically identical to the cumulative word count, we want to emphasize that while word count is often thought of as a static measure, word-time flows as a reader goes through a text.

However, research on word usage change in word-time is scant. To our knowledge, there has yet to be a quantitative study on how word usage changes in word-time across different texts. We know from the prior studies mentioned earlier that word usage changes over the length of a book in a general manner. Just as a larger narrative can be broken down into sub-narratives (Wallace, 2012), can we find meaningful changes in word usage within sections of a book? How do we characterize these changes?

While there are a number of ways to quantify word usage, we are particularly interested in how essential meaning changes over word-time. Since the 1950s, there have been efforts to distill the meaning of words into a few numbers. The valence–arousal–dominance (VAD) framework (Osgood et al., 1957) has been one of the most widely used systems in quantifying meaning through semantic differentials. This framework was developed using factor analysis of word scores provided by human annotators for a small set of words, which resulted in the three dimensions of valence, arousal, and dominance. However, efforts to create large VAD lexicons using human annotators (Mohammad, 2018; Warriner et al., 2013) have shown that valence, arousal, and dominance are linearly correlated when a larger set of words is used. By transforming the National Research Council Canada (NRC) VAD lexicon—the largest VAD lexicon published with more than 20,000 curated words (Mohammad, 2018)—via singular value decomposition, Dodds et al. (2021) showed that the VAD framework can be collapsed onto two linearly independent dimensions which align with the concepts of “power” and “danger”, with the third linearly independent dimension (“structure”) being less relevant in real-world corpora than the former two. Figure 1 shows representative words in power-danger space for Terry Pratchett's Discworld series. Real-world corpora also exhibit a bias toward low-danger words, which parallels the positivity bias in the language (Dodds et al., 2015). Valence scores are highly correlated with happiness scores (Dodds et al., 2021), which have been used to quantitatively uncover emotional arcs in stories (Reagan et al., 2016). Thus, we expect power-danger scores, both having a linear dependence on valence (see Supplementary Information), to also be suitable markers in quantifying structure in text and to give more information than using happiness scores alone, especially since the power-danger lexicon contains more words and would yield higher word coverage.

An illustration of how a word usage marker, such as the danger score, changes over the course of a text is given in Fig. 2A–D. Each time series was constructed by splitting the text into windows of size  $N_w$  words that skip every  $N_s$  words; in previous studies, time series for all books were constructed using the same value for  $N_w$ , which range from a few hundred to a few thousand, with either  $N_s$  or the number of windows fixed. Each window corresponds to a point in the time series and is characterized by the mean score of the words it contains. Using a larger window size smooths out the time series, sometimes revealing the general shape, such as the steady increase in danger scores in “The Strange Case of Dr. Jekyll and Mr. Hyde” by Robert Louis Stevenson, or an up-down pattern as in “The Winter's Tale” by William Shakespeare. Longer books, however, tend to retain fluctuations at the same window sizes. If books of different lengths are compared by the fraction of the book covered, which we define as the *normalized word-time* (Fig. 2E), it would seem that word usage changes more steadily for shorter books than

~ power-danger ousiogram for the Discworld series ~



**Fig. 1** An ‘ousiogram’ (Dodds et al., 2021) displaying power and danger scores for a subset of 14,499 unique words appearing in Terry Pratchett’s 41-book Discworld series. The example words overlaid map the indicated eight major directions of the compass of essential meaning. The full set of 20,006 words with power and danger scores is derived from the NRC VAD lexicon (Mohammad, 2018). These 14,499 unique words collectively account for 1,140,027 total words (types versus tokens). The histogram’s color map indicates the frequency of usage. The histogram, along with the marginal distributions for power and danger (right and top) presents the same safety bias observed across disparate corpora (Dodds et al., 2021).

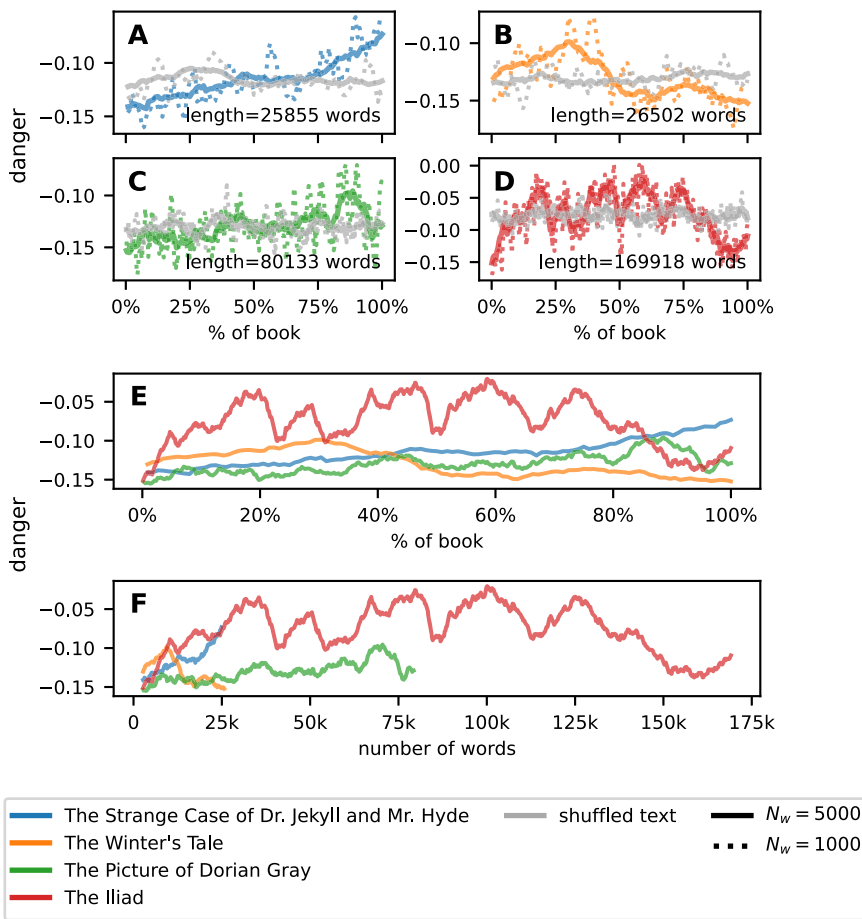
longer books. However, if we compare the time series in (raw) word-time, so that the time series are of different lengths (Fig. 2F), the shorter time series appear comparable to sections of the longer time series. Thus, changes in word usage may be related to the length of a text, with longer texts having a similar structure to a concatenation of shorter texts. In the context of narratives (as our examples in Fig. 2 are), this suggests the possibility that long narratives may be composed of shorter ones, each with its own arc, that function as the basic unit of the story. Some of the fluctuations found in the time series may not be unwanted noise, but rather a measure of the lengths of these basic units. Our aim is to characterize these fluctuations for various texts and to relate them to the properties of the texts themselves.

With this objective, we cannot use large window sizes, since they run the risk of smoothing out potentially relevant fluctuations. However, we must also isolate the fluctuations that arise from noise, something we will likely see with smaller window sizes. Since we are interested in how word usage changes over the course of a text, we can compare the time series to that obtained from a shuffled version of the text, which contains the same words but orders them randomly. Using this as a reference isolates the effect of word order, and also avoids making assumptions on the nature of contaminating noise. As expected, time series obtained from shuffled texts are flatter, with smaller fluctuations than those found in the original text (Fig. 2A–D).

To extract and characterize fluctuations in the time series at different scales, we use empirical mode decomposition (EMD), a technique that factors a signal into a sum of *internal mode functions* (IMF), each of which is a mean-zero oscillatory time series with frequency and amplitude modulations, and a non-oscillating trend (Huang et al., 1998). While it is similar to wavelet decomposition in terms of its objective, EMD is data-adaptive, requiring almost no critical input from the user other than the raw time series itself, and is also well-suited for both nonstationarity and nonlinearity in time series. A more detailed explanation of EMD is given in the Supplementary Information.

Figure 3 shows the result of performing ensemble empirical decomposition (EEMD), an EMD variant that is more robust to noise (Wu and Huang, 2009), on a time series obtained from “The Iliad”. While the original time series, derived from small, non-overlapping windows, contains significant noise, EEMD is able to separate the signal into components of different characteristic frequencies. Further, we see that a partial reconstruction of the time series, obtained by summing the low-frequency IMFs, can replicate the time series obtained with larger window sizes, producing a denoised version.

We mentioned earlier that the time series for shuffled text is flatter than that of the original text (Fig. 2A–D), indicating that the IMFs of original and shuffled text may differ in their variances. An illustration of how the variance changes as the IMF order increases for both the original and shuffled texts is given in



**Fig. 2 Time series for books of different lengths in raw and normalized word-time. A–D** These plots illustrate the effect of window sizes, shuffling, and book length on how danger scores change over the course of a book. Using larger window sizes (solid vs. dotted) results in smoother curves, although longer books tend to retain fluctuations in the time series for the same window size (A, B vs. C, D). Shuffling the text results in time series that are flatter than the original (gray). **E** Comparing books using normalized word-time shows more fluctuations for longer books than shorter ones, but **F** plotting the time series in raw word-time shows similarities between shorter books and sections of longer books. A skip size of  $N_s = 200$  words was used for all plots shown above.

Fig. 4. The variance of the IMFs in the shuffled texts generally decreases as the IMF order increases, similar to the observation for fractional Gaussian noise (Flandrin et al., 2004; Wu and Huang, 2004). However, this is not always true for the original texts. For example, “The Iliad” shows a clear example of a book that differs in variance from the shuffled text from an IMF order below the trend, and continues to do so until the trend level. On the other hand, “The Picture of Dorian Gray” clearly differs in the variance at an IMF order below the trend, but does not always do so for higher IMF orders. In some books, such as “The Strange Case of Dr. Jekyll and Mr. Hyde”, the original and shuffled versions only differ in the variance at the trend level.

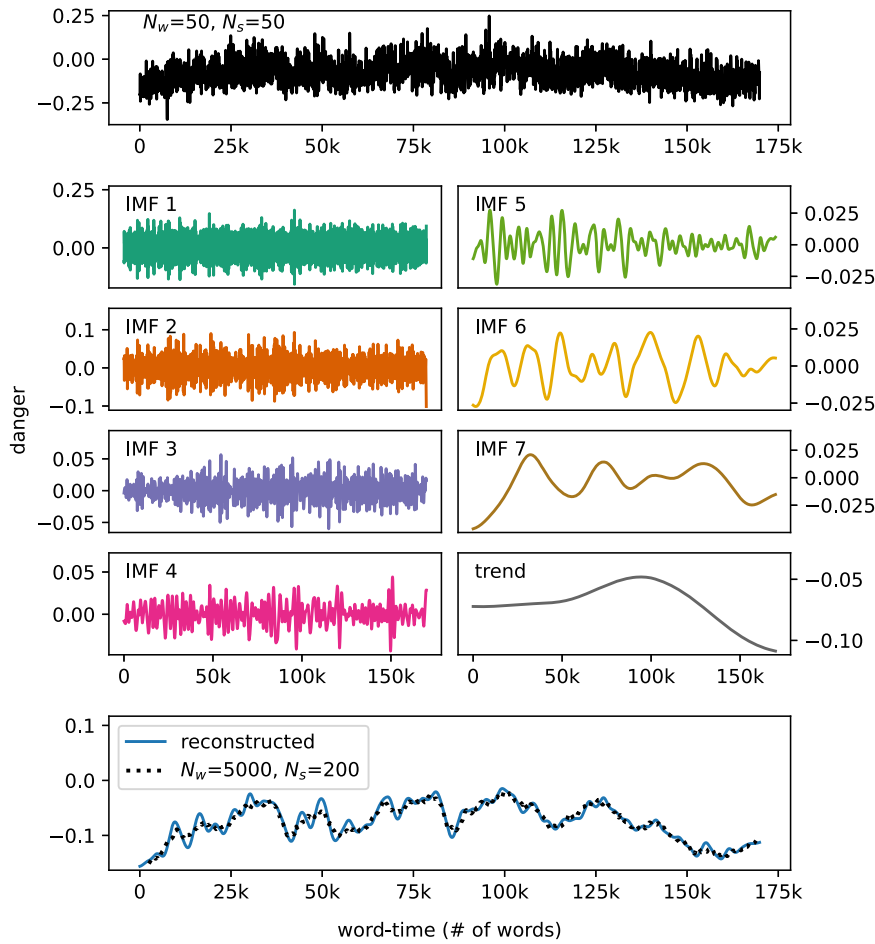
To identify the cutoff IMF order for each book, we compare IMFs of the original text from those of different realizations of shuffled text. The lowest IMF order at which the variance is higher than expected from the shuffled version is considered the cutoff at which word order becomes relevant. Those where the cutoff order is an IMF (i.e., not the non-oscillating trend) are considered to have relevant fluctuations on top of the trend, while those that do not are considered to be trend-only. To compare the variance, we use the method used by Wu and Huang (2004) and Flandrin et al. (2004), where the variances of the IMFs of the target series are rescaled so that the variances of the first IMFs of the target (original text) and reference (shuffled text) are comparable. The assumption here is that the first IMF is noise; as

we are using non-overlapping windows of size  $N_w = 50$ , it is very difficult for a coherent narrative to be present at this scale, and the first IMF will most likely pick up noise due to the small window size. Pairwise comparison of each IMF for the original and shuffled versions shows that the IMFs have similar periods for the first IMF, supporting this assumption. We also verify that the periods are comparable up to the cutoff IMF (Fig. S1). While rescaling to the median of the first IMF is a reasonable choice, we also examine how rescaling to the 1st percentile of the first IMF or having no rescaling affects the results.

While our study has been motivated by both qualitative and quantitative studies on narrative, we wish to study how word usage changes in word-time for a broader set of texts, both fiction, and nonfiction, and discuss results for different categories as given by their Library of Congress Classification class and subclass labels.

## Methodology

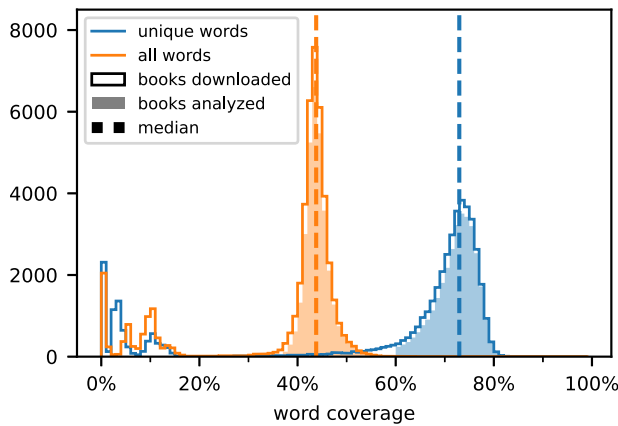
**Data preprocessing.** We downloaded more than 45,000 books from Project Gutenberg (Project Gutenberg, n.d.), an online repository of books in the public domain. The Gutenberg headers were removed using code from the Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2020). Contractions, when unambiguous, were replaced with their expanded versions (e.g., “n’t” to “not”); if ambiguous, they were deleted, similar to what



**Fig. 3 The decomposition of the danger time series of "The Iliad" using EEMD.** The raw time series was obtained using non-overlapping windows with  $N_w = 50, N_s = 50$  (top panel), and the IMFs and the trend obtained from EEMD are shown. The lowermost panel shows the sum of IMFs 5–7 and the trend ("reconstructed"), superimposed with the danger time series obtained using larger overlapping windows ( $N_w = 5000, N_s = 200$ ). Note that the partial reconstruction from the time series obtained using smaller windows is very similar to the raw time series obtained using larger, overlapping windows.



**Fig. 4 A comparison of the variances of the IMFs for the original text and 100 different realizations of shuffled text for three books in the dataset.** Lower-order IMFs are more likely to correspond to noise, while higher-order IMFs are more likely to contain relevant information. Note that while technically the trend is not an IMF, for comparison purposes, the trend is included in this plot as the highest IMF order found for a given time series.



**Fig. 5 Word coverage of the PDS lexicon.** The blue histograms correspond to the unique word coverage, while the orange histograms correspond to the word coverage allowing for word repetition. We show the histograms for all the books downloaded in our dataset (empty bars) as well as those we analyzed (shaded bars), and the dashed lines show the median word coverage of the books analyzed. The books included in our analysis do not include duplicate titles and must have (1) a unique word coverage of at least 60%, (2) at least one lexicon word in every window, and (3) intrinsic mode functions (IMFs) that can be successfully computed.

was done in Fudolig et al. (2022). The remaining text was then converted to lowercase and tokenized using whitespace as separators, disregarding words that contain non-word characters and digits, and ignoring punctuation marks. This converts the text into a sequence of words.

We then examine the word coverage of the power-danger lexicon. While the original NRC-VAD lexicon (Mohammad, 2018), from which the danger and power scores were derived, contained around 20,000 words, we expanded this to include noun plurals and conjugated forms not in the lexicon. Scores of the base forms of the verbs and nouns were used as the scores of their conjugated versions, expanding the lexicon to 32,721 words (the lexicon is available at <https://doi.org/10.5281/zenodo.7816312>). We only consider books with a 60% unique word coverage, which consist almost exclusively of English text and cover 93% of all English books in the downloaded set. The median unique word coverage of this subset of books is 73%. Filtering for word coverage, removing books with duplicate titles, as well as requiring that the time series must have at least one lexicon word and that the ensemble empirical mode decomposition (EEMD; see the section “Characterizing relevant fluctuations”) can be successfully computed (i.e., the EEMD decomposes up to the trend level, such that the mean of the sum of the EEMD results is within 10% of the mean of the original time series), leaves us with 31,690 books (Fig. 5).

**Constructing the time series from text.** We construct the danger and power time series by segmenting the sequence of words into non-overlapping windows of size  $N_w = 50$ , each of which corresponding to a point in the time series. In each window, we take each word  $w_i$  with score  $s_i$  in the lexicon that occurs  $n_i$  times in the window. If there are  $m$  unique words in the window that are in the lexicon, then the score for the window is

$$s_w = \frac{\sum_{i=1}^m n_i s_i}{\sum_{i=1}^m n_i} \quad (1)$$

The reference time series are constructed by shuffling the tokenized version of the text, performing the windowing technique, and recomputing the average scores for each window.

This assures that the time series of both the target (original text) and the reference (shuffled text) are of the same length.

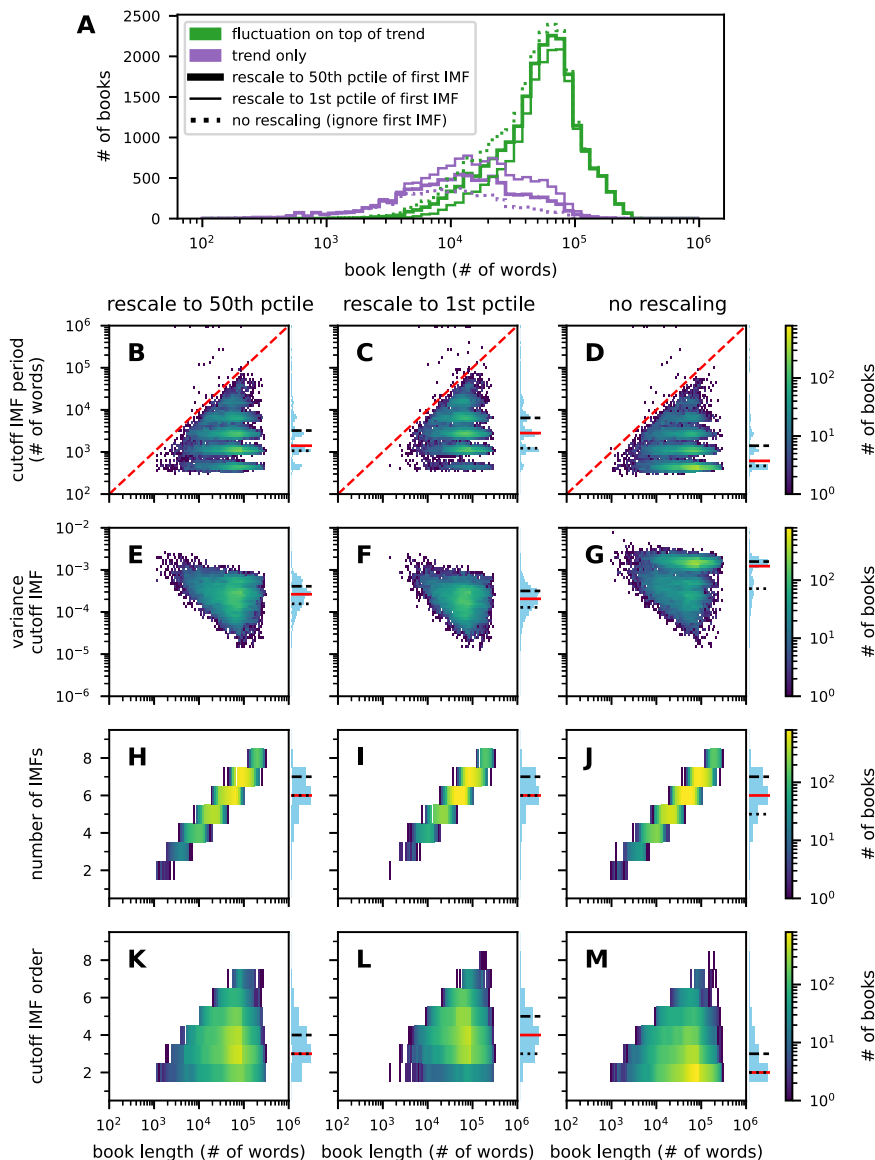
**Characterizing relevant fluctuations.** We use ensemble empirical mode decomposition (EEMD) (Wu and Huang, 2009) with an ensemble size of 100 to obtain the internal mode functions (IMFs) for the time series corresponding to the original text. Each time series in the ensemble is the sum of the raw time series and white noise with a standard deviation of  $0.2\sigma$ , where  $\sigma$  is the standard deviation of the raw time series, as suggested by Wu and Huang (2009). As our reference time series, we use 100 different shuffled versions of the original text. For each of the reference time series, we use basic empirical mode decomposition (EMD) to obtain the IMFs. We use code from the Python package `emd` (Quinn et al., 2021) to perform all EMD-related calculations.

We then rescale the variances of the IMFs of the time series as suggested in (Flandrin et al. 2005; Wu and Huang, 2004). The variance for an IMF that has zero mean by definition is  $\text{Var} = \sum_i x_i^2 / N$ , where  $N$  is the length of the time series and  $x_i$  is the value at word-time point  $i$ . If the variance of the IMF order  $i$  in the target (original text) is  $\text{Var}_{T,i}$  and  $\text{Var}_{R,1}$  is the representative value of the first IMF of the reference (shuffled text), then the rescaled variances are  $\text{Var}'_{T,i} = \text{Var}_{T,i} \frac{\text{Var}_{R,1}}{\text{Var}_{T,1}}$ . In the log-scale plot given in Fig. 4, this is equivalent to shifting the variance curve of the original text up or down so that the rescaled first IMF of the original text is equal to a representative value of the first IMFs of the shuffled texts ( $\text{Var}'_{T,1} = \text{Var}_{R,1}$ ). Three different representative values were considered based on the distribution of the first IMFs of the reference time series: the median, the 1st percentile, and the variance of the first IMF of the target (no rescaling). The lowest IMF order at which the rescaled variance is higher than the 99th percentile of the variances for shuffled text is considered as the cutoff IMF order.

Once the cutoff IMF order is identified, the corresponding period is computed from the center of the frequency bin with the highest energy as obtained using the Hilbert–Huang transform (HHT). Since we want to count periods in the unit of number of words, we compute for the HHT setting the sampling rate at  $N_s^{-1} \text{word}^{-1}$ , where  $N_s = 50$  is the skip size in the windowing procedure. We use logarithmically spaced frequency bins, spanning from  $10^{-6} \text{word}^{-1}$  to  $1 \text{word}^{-1}$ , resulting in a range of period values between 1 to  $10^6$  words, chosen because none of the texts examined exceed  $10^6$  words in length. The choice of logarithmic spacing is motivated in part by how EMD on white noise performs like a dyadic filter, with IMF frequencies decreasing by roughly a factor of 2 for every order. Further, in our preliminary analysis for select books, we find that logarithmic spacing provides an adequate representation of the spectra, especially since the IMF frequencies and periods span orders of magnitude. We emphasize that obtaining a characteristic value for the *period* of the IMF is independent of the method used to extract the cutoff IMF order discussed earlier, and that the bins used for the HHT are the same across all texts.

## Results

We analyze more than 30,000 books from Project Gutenberg that passed our selection criteria. The selected books are almost exclusively in English, with at least 60% of the unique words in each book included in the lexicon. Around 60% of the books are in the “Language and Literature” Library of Congress Classification (LCC) code (class label “P”), while the remaining are spread out among the various LCC class labels, with “World History” (class label “D”) as the next largest category of books in the



**Fig. 6 Characterizing relevant fluctuations in the danger time series of books in the corpus.** **A** These histograms show the number of books that have fluctuations on top of the trend (green) and those that do not (purple). The different line widths and line styles correspond to the different rescaling factors: solid thick lines for using the median of the first IMF, solid thin lines for using the 1st percentile of the first IMF, and dotted lines for no rescaling. **B-D** These are heatmaps that show the relationship between the period of the cutoff IMFs (as the number of words) and the book length for the various rescaling factors (shown above each plot). We can see that most of the points fall below the 45-degree line (red dashed line), indicating that the cutoff IMF period is less than the book length for the vast majority of books. The histograms for the cutoff IMF period are shown on the right side of each plot, with the 25th, 50th, and 75th percentiles shown in dotted black, solid red, and dashed black lines, respectively. The rest of the figures are similar to (**B-D**), but for different quantities: variance (**E-G**), number of IMFs (**H-J**; excludes the trend), and cutoff IMF order, with the first IMF counted as 1 (**K-M**).

dataset (around 8% of the books). While we performed the analysis for both danger and power scores, the results are similar for both, and we only discuss the danger scores in the main text. The results for power scores are included in the Supplementary Information.

**Cutoff IMF orders.** Figure 6A shows histograms of the number of words found in books that have a cutoff IMF order below the trend, and those that do not. While the choice of the rescaling factor influences the stringency of the cutoff criteria, they also offer insight into the robustness of the results.

Rescaling to the 1st percentile of the first IMF requires a higher threshold to differentiate the IMFs, and thus more books are classified as trend-only. For those classified as

having fluctuations on top of the trend, the cutoff periods obtained may be higher. On the other hand, no rescaling makes it more likely to see differences in the lower IMF orders, resulting in lower predicted cutoff periods. We note that when no rescaling is performed, the cutoff IMF order for many of the books was the first IMF, which does not make much sense given that the window size of 50 words is relatively small and that the first IMF only corresponds to a period of around 100 words. We also know from the IMF decomposition of fractional Gaussian noise (Flandrin et al., 2005; Wu and Huang, 2004) that the first IMF has a different behavior compared to the higher-order IMFs. Thus, for purposes of comparison, we disregard the first IMF in obtaining the cutoff IMF order when no rescaling is applied.

The predicted cutoff IMF order, including the case when it does not exist, differs for the majority of the books examined. However, we find some general results that hold regardless of the rescaling factor used. For instance, we find that longer books tend to have relevant fluctuations on top of the trend (i.e., have a cutoff IMF order that is not the trend) while shorter books do not (Fig. 6A), especially for books with <3000 words or greater than 100,000 words. Books with word counts in between these values may or may not have relevant fluctuations on top of the trend and the choice of the rescaling factor may influence the prediction for the cutoff IMF order. When rescaling by the 50th percentile of the first IMF, the 25th to 75th percentiles range from roughly 1000–3200 words; when rescaling by the 1st percentile of the first IMF, this changes to around 1200 to 6400 words; and when no rescaling is used, this changes to around 500–1400 words. In the case when relevant fluctuations on top of the trend are found, no clear relation between the cutoff IMF period and the book length is observed. These can be seen in the heat maps and the corresponding histograms in Fig. 6B–D. We also note that, with very few exceptions, the cutoff period is less than the book length. The striated pattern in the heatmaps for the cutoff periods is due to the discreteness of the EEMD, which is observed even in white noise (Wu and Huang, 2004), where the EEMD acts like a dyadic filter. While these results were generated using all the books in the corpus, both fiction and nonfiction, we also analyze the different book classifications in the next subsection and include more details in the Supplementary Information.

Similarly, the raw variances of the cutoff IMFs also do not exhibit any relationship with the book length (Fig. 6E–G). The range of the variance values is also wider in comparison to that observed for the first IMF (Fig. S1E–F), indicating that while the first IMF likely corresponds to noise due to the windowing technique used for all of the books, the books generally start to differ from each other at higher IMF orders. Further, while the number of IMFs found by EEMD for the original time series increases with the book length (Fig. 6H–J), the cutoff IMF order itself shows no such correlation (Fig. 6K–M).

We find similar observations from our analysis of the power time series. Shorter books are more likely to be trend-only, while longer books have relevant fluctuations above the general trend. While the cutoff periods are not identical to those found in danger scores, they are of the same order of magnitude (see Supplementary Information, Fig. S3, Table S1).

**Relationship to book content.** The Gutenberg corpus assigns books to their Library of Congress Classification (LCC) subclass labels (e.g., PS for American Literature). While it is possible for a book to have more than one LCC subclass label, around 95% of the books we examined only had 1 label. The top 5 subclass labels with the most books in the dataset are American literature (PS), English literature (PR), Fiction and juvenile belles lettres (PZ), Periodicals (AP), and French/Italian/Spanish/Portuguese literature (PQ). We also looked at the class labels rather than the subclass labels (i.e., the first letter of the subclass labels). Around 60% of the class labels are for Language and Literature (P), while World History and History of Europe, Asia, Africa, Australia, New Zealand, etc. (D); Philosophy, Psychology and Religion (B); General Works (A); and History of America (E) round up the top 5 class labels. While there are differences in the medians for the cutoff periods and variances across subclass and class labels, the spread of both the cutoff IMF period and variance are comparable across the top 5 class and subclass labels. This indicates that the breadth of the LCC system at the class or subclass levels obscures the differences that may arise from smaller groups within these categories (Fig. 7).

On the other hand, we find that using very specific filters on the title of the book yields more insightful results (Fig. 8). For books with a word beginning with “poem” in the title, those with a cutoff IMF order below the trend exhibit shorter cutoff IMF periods and markedly higher cutoff IMF variances compared to the rest of the dataset, which is consistent with poems being typically short, emotional and compact. On the other hand, books with titles containing a word beginning with “manual” exhibit a lower median cutoff IMF variance. This may be because books in this category (“A manual of clinical diagnosis”, “The ladies’ book of etiquette, and manual of politeness: a complete handbook for the use of the lady in polite society”, “The skillful cook: a practical manual of modern experience,” etc.) tend to be instructional and uniform in terms of topic and mood. Books with words beginning with “play” have a higher median cutoff IMF period and lower cutoff IMF variance than books without. Results for keywords such as “collection”, “short stor” (includes both “short story” and “short stories”), “report” and “essay” are more sensitive to the choice of rescaling factor.

For power time series, while the values for the cutoff IMF periods and variances obtained are different, we also find similar overlap across different LCC class and subclass labels. We also find differences across different books depending on words in their titles (see Supplementary Information, Figs. S4 and S5).

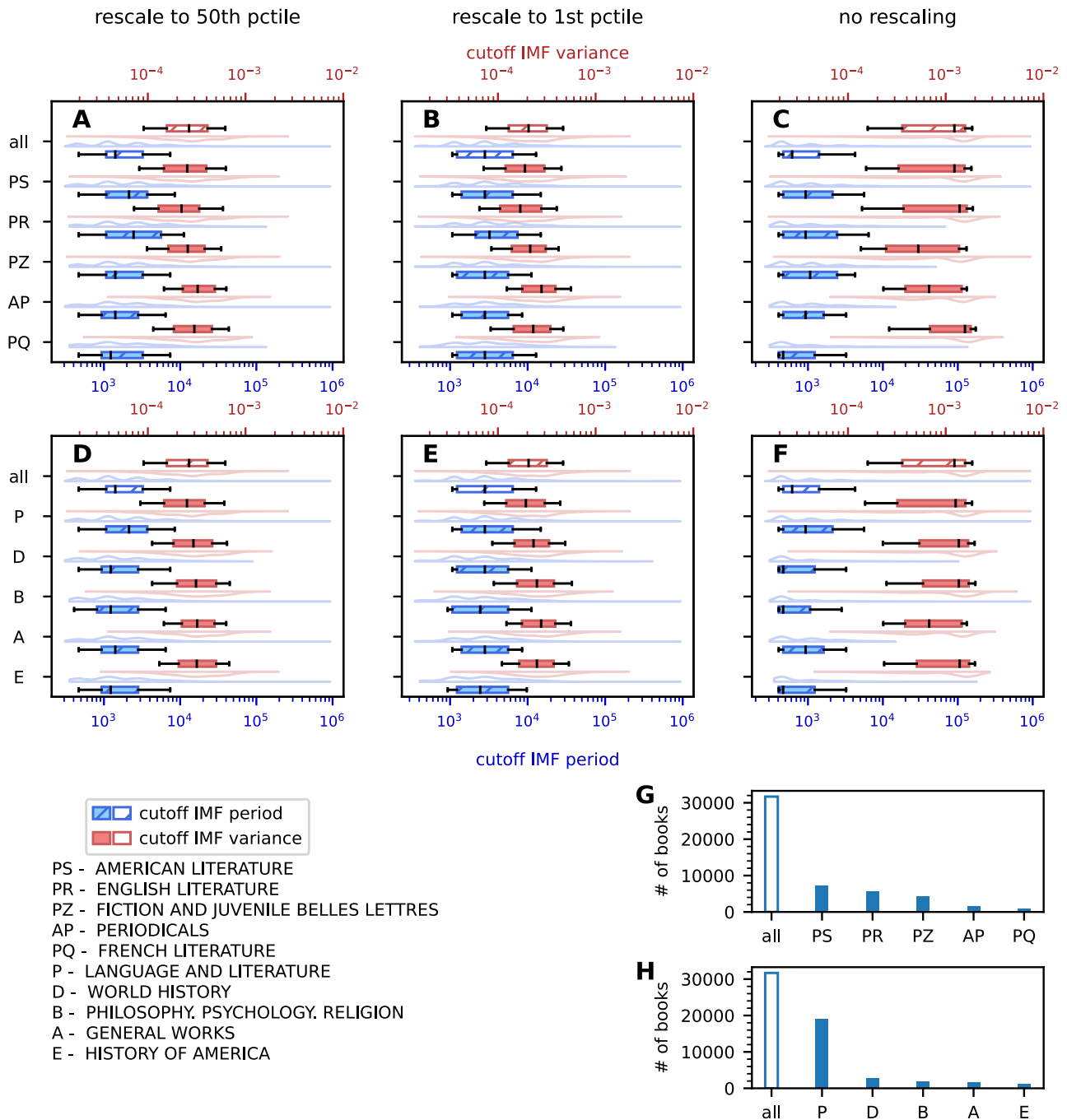
## Discussion

**Summary of results.** We examine the fluctuations in the danger and power dimensions in ouisiometrics, a reinterpretation of the valence–arousal–dominance framework, for more than 30,000 English books in Project Gutenberg. Changes in word usage across text are analyzed by segmenting the text into windows and converting it to a time series. While window size has conventionally been preset independent of the book length, we find that large window sizes remove oscillations in word usage in shorter books but retain them in longer books. However, a visual examination of the time series in word-time, which we define as the number of words seen up to the present moment in a book, seems to indicate that the time series for shorter text is similar to subsections of a longer text. This suggests that in the ouisiometric sense, longer texts are similar in structure to a concatenation of shorter texts, not unlike a novel broken into chapters. We verify this by obtaining quantitative estimates of the periods of the relevant fluctuations in the time series.

We extract the different scales of fluctuations in the time series obtained from text using empirical mode decomposition (EMD). EMD decomposes a time series into a non-oscillatory trend and a sequence of oscillatory intrinsic mode functions (IMFs) that differ in their characteristic frequencies. It allows for both nonlinearity and nonstationarity and gives an intuitive and data-driven understanding of the underlying fluctuations in a given time series.

For each book, we derive the danger and power time series for both the original text as well as an ensemble of shuffled versions. By comparing the variances of the resulting IMFs of the time series obtained from the original text to those of the shuffled texts, we find that shorter books tend to exhibit only a non-oscillatory trend while longer books tend to exhibit relevant fluctuations with periods around the order of a few thousand words, shorter than the length of the book. The period and variance of the relevant fluctuations do not depend on the book length, despite the number of IMFs increasing with the book length. Segregation of books by the Library of Congress Classification class or subclass codes does not result in well-defined groups that show variation in the periods and variances of the relevant fluctuations. However, we observed differences when we applied very specific



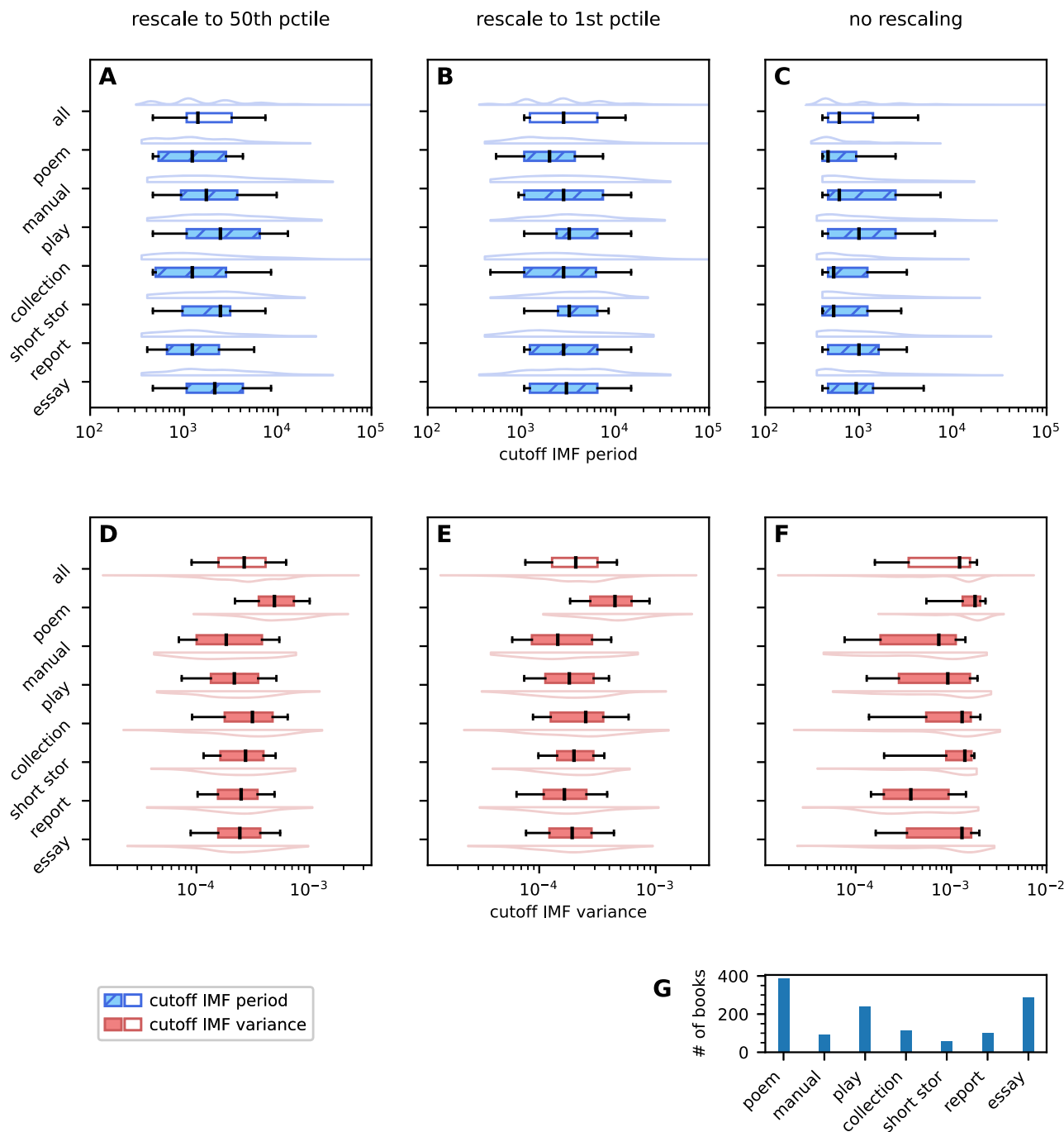


**Fig. 7** Periods and variances of the cutoff IMF in the danger time series for the top 5 LCC subclass and class labels with the most number of books in the dataset examined. **A-F** These are the boxplots for the periods (see x-axis at the bottom for scale) and the variances (see x-axis at the top for scale) of the cutoff IMF for books in the entire dataset ("all") as well as that for each of the top 5 LCC subclass and class labels by size. The black line inside each box is the median, the box ranges from the 25th to 75th percentiles, and the whiskers extend from the 9th percentile to the 91st percentile. Violin plots are also included for more detail. **G** and **H** show the number of books analyzed in the dataset ("all") and the top five subclass and class labels, including trend-only books.

filters, such as words in a title. We infer that what characterizes the scale of a book's relevant fluctuations is neither its general topic nor length, but rather more specific aspects such as its structure and content (e.g., poems vs. manuals).

The impact of our study is mainly on two fronts: (1) a quantitative analysis of the structure of texts as a function of word-time that is consistent with how longer texts are divided into meaningful sections, and (2) a method for denoising time series obtained from texts without resorting to using arbitrarily large window sizes. We discuss these in the following paragraphs.

**Word counts and book types.** Segmenting longer texts into shorter, self-contained sections has long been acknowledged in the publishing world, as well as by researchers in the literature and computational domain (Reagan et al., 2016; Wallace, 2012). While it is theoretically possible to write a basic narrative in a few words (e.g., "For sale, baby shoes, never worn.") or a hundred thousand, balancing the flexibility of longer texts with the constraints imposed by reader engagement and publication costs has given writers and editors rules of thumb for word counts.



**Fig. 8** Periods and variances in the danger time series for books with a word in the title if they have a cutoff IMF order below the trend. Note that since the number of books for each word is much less than the total number of books examined, the boxplots for books not containing a given word in the title will be almost identical to the boxplots for the entire dataset ("all"). **A-C** show the boxplots for the cutoff IMF period, while **D-F** show the boxplots for the cutoff IMF variance for different rescaling factors. The black line inside each box is the median, the box ranges from the 25th to 75th percentiles, and the whiskers extend from the 9th percentile to the 91st percentile. Violin plots are also included for more detail. **G** shows the number of books with the given keywords in the title, including trend-only books.

Short stories and novellas are characterized by a focus on a single central conflict (MasterClass, 2021) or a single chain of events (Baldick, 2015). In contrast, the longer word count of novels (>40,000 words Science Fiction & Fantasy Writers of America, 2020; World Science Fiction Society, 2022) allows for a fuller development of its characters and themes (Baldick, 2015) through the use of chapters, each of which is similar to a short story. These are consistent with our results: trend-only books begin to decrease in number above a word count of 10,000, while

the number of books with relevant fluctuations above the trend continues to increase, reaching its peak for books with 50,000–100,000 words (Fig. 6). Further, the cutoff IMF periods are in the order of a few thousand words, comparable to the length of chapters, which are typically 1500–5000 words long (Bingham, 2020, "How long should a chapter be", 2017).

While these are editorial guidelines for stories, we find that word usage in the ousiometric sense, through the power and danger time series, shows segmentation for longer books across

several categories, even those that do not necessarily qualify as literary narratives (Piper et al., 2021) (see Supplementary Information for a comparison between literature and non-literature books, Table S1, Figs. S6–S9). Prior quantitative studies have shown that word usage patterns change over the course of the text, not just for literary narratives but for other types of text (e.g., academic papers), albeit with different signatures (Boyd et al., 2020; Pechenick et al., 2015). Segmentation in non-literary books is not uncommon as evidenced by the widespread use of chapters and sections, and is supported by our results.

Another result we wish to highlight is that the period of relevant fluctuations is independent of the book length. As this period can be interpreted as the word count of meaningful segments, this suggests the existence of a basic unit of the text of some characteristic length that serves as a building block to construct longer texts. We suspect this may relate to the rate at which humans can process textual information, although testing this hypothesis is outside the scope of our work.

**Data-adaptive denoising of text-derived time series.** On the quantitative side, the method we used in this paper allows us to smooth out fluctuations in time series derived from the text of various lengths. While using large window sizes reduces noise, it is unclear if it also inadvertently smooths out relevant information, such as fluctuations associated with subplots. By performing partial reconstruction using the relevant IMF orders, we have a data-driven denoising approach that works for both short and long texts.

As our method relies heavily on empirical mode decomposition, it also carries the same limitations. Although EMD will ideally construct different modes for fluctuations of sufficiently different frequencies, mode mixing may occur due to signal intermittency. Using ensemble EMD (EEMD) mitigates this problem, but does not ensure that it will remove mode mixing in all cases. We also note that we compare the original and shuffled texts only in the variance of their IMFs. While we verified that the pairwise comparison of IMFs results in comparable IMF periods for the original and shuffled text up to the cutoff IMF order, we only defined a difference in IMFs in terms of their variance. While this method has been proposed for finding the appropriate cutoff IMF order in using EMD for denoising (Flandrin et al., 2005; Wu and Huang, 2004), it will miss any other difference that is not associated with the variance. We also considered using the probability density function to compare IMFs; however, this method failed to produce accurate results in synthetic data, while the variance comparison method performed well in all the tests we performed. While our general observations are robust to the choice of parameters in the variance comparison method, different results may be obtained for a particular book depending on the parameters used. Thus, while our method can extract general trends on relevant fluctuations across a corpus, sensitivity analysis must be performed when results are to be obtained for a particular book.

**Future work.** As discussed earlier, our work opens up avenues to investigate building blocks of text in books; narratives would particularly be of interest. In our study, we only look at the minimum word count at which word order becomes relevant in the ousiometric sense. However, in addition to the cutoff IMF order from which this word count was obtained, the denoised ousiometric time series also includes contributions from higher IMF orders. It would also be interesting to see whether there is a hierarchical structure in text segmentation.

We have studied the scale of fluctuations within texts in their danger and power time series. While danger and power are orthogonal to each other and produce similar temporal results, we

did not look at how they work together as a book progresses. Similar to the work by Toubia et al. (2021), we hope to do a spatial analysis in power-danger space, specifically on the path taken by the narrative. Other possible avenues for future work include expanding our corpus to include various texts, such as screenplays and movies, as well as comparing different versions of a book, such as the first draft and the final published version.

### Data availability

The source data is publicly available from Project Gutenberg (Project Gutenberg, n.d.). Instructions on how to generate and process the data used in the study are available at <https://doi.org/10.5281/zenodo.7816312>.

Received: 8 September 2022; Accepted: 13 April 2023;

Published online: 29 April 2023

### References

- Ayenu-Prah A, Attoh-Okine N (2010) A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition. *Adv Adapt Data Anal* 2:1–24
- Baldick C (2015) *The Oxford Dictionary of Literary Terms*. Oxford University Press
- Bingham H (2020) How long should a chapter be? <https://jerichowriters.com/how-long-should-a-chapter-be/>
- Boudraa A, Cexus J, Benramdane S, Beghdadi A (2007) Noise filtering using empirical mode decomposition. In: 2007 9th international symposium on signal processing and its applications, Sharjah, United Arab Emirates. IEEE, pp. 1–4
- Boudraa A-O, Cexus J-C (2007) EMD-based signal filtering. *IEEE Trans Instrum Meas* 56:2196–2202
- Boyd RL, Blackburn KG, Pennebaker JW (2020) The narrative arc: revealing core narrative structures through text analysis. *Sci Adv* 6:eaba2196
- Brown S, Tu C (2020) The shapes of stories: a “resonator” model of plot structure. *Front Narrat Stud* 6:259–288
- Corral Á, Boleda G, Ferrer-i-Cancho R (2015) Zipf’s Law for word frequencies: word forms versus Lemmas in long texts. *PLoS ONE* 10:e0129031
- Dodds PS, Alshaabi T, Fudolig MI, Zimmerman JW, Lovato J, Beaulieu S, Minot JR, Arnold, MV, Reagan AJ, Danforth CM (2021) Ousiometrics and telegonomics: the essence of meaning conforms to a two-dimensional powerful-weak and dangerous-safe framework with diverse corpora presenting a safety bias. *arXiv:2110.06847 [physics]*
- Dodds PS et al (2015) Human language reveals a universal positivity bias. *Proc Natl Acad Sci USA* 112:2389–2394
- Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLoS ONE* 6:e26752
- Dodds PS et al (2020) Allotaxonomy and rank-turbulence divergence: a universal instrument for comparing complex systems. *arXiv:2002.09770 [physics]*
- Flandrin P, Gonçalves P, Rilling G (2005) Emd equivalent filter banks, from interpretation to applications. In: Huang NE, Shen SSP (eds) *Hilbert–Huang transform and its applications*, vol 5 of interdisciplinary mathematical sciences. World Scientific, pp. 57–74
- Flandrin P, Rilling G, Goncalves P (2004) Empirical mode decomposition as a filter bank. *IEEE Signal Process Lett* 11:112–114
- Freytag G (1900) *Freytag’s Technique of the drama: an exposition of dramatic composition and art*. An authorized translation from the 6th German ed., 3rd edn. Scott, Foresman, Chicago
- Fudolig MI, Alshaabi T, Arnold MV, Danforth CM, Dodds PS (2022) Sentiment and structure in word co-occurrence networks on Twitter. *Appl Netw Sci* 7:1–27
- Gao J, Jockers ML, Laudun J, Tangherlini T (2016) A multiscale theory for the dynamical evolution of sentiment in novels. In: 2016 International conference on Behavioral, Economic and Socio-cultural Computing (BESC). pp. 1–4
- Genette G (1983) *Narrative discourse: an essay in method*. Cornell University Press
- Gerlach M, Font-Clos F (2020) A standardized Project Gutenberg Corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* 22:126
- How Long Should A Chapter Be? (2017) *The master guide to chapter length*. <https://blog.reedsy.com/how-long-should-a-chapter-be/>

- Huang NE et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc R Soc Lond Ser A* 454:903–995
- Komaty A, Boudraa A-O, Augier B, Daré-Emzivat D (2014) EMD-based filtering using similarity measure between probability density functions of IMFs. *IEEE Trans Instrum Meas* 63:27–34
- MasterClass (2021) Learn the differences between Novelettes, Novellas, and Novels. <https://www.masterclass.com/articles/learn-the-differences-between-novelettes-novellas-and-novels>
- Mohammad S (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics, vol 1: Long papers. Association for Computational Linguistics, Melbourne, Australia, pp. 174–184
- Moretti F (2013) Distant reading. Verso, New York
- Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning. University of Illinois Press
- Pechenick EA, Danforth CM, Dodds PS (2015) Characterizing the Google Books Corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10:e0137041
- Phelan J, Rabinowitz P (2012) Time, plot, progression. In: Narrative theory: core concepts and critical debates. Ohio University Press, pp. 57–65
- Piper A, So RJ, Bamman D (2021) Narrative theory for computational narrative understanding. In: Proceedings of the 2021 conference on empirical methods in natural language processing, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 298–311
- Project Gutenberg (n.d.) <https://www.gutenberg.org/>
- Quinn A, Lopes-dos Santos V, Dupret D, Nobre A, Woolrich M (2021) EMD: empirical mode decomposition and Hilbert–Huang spectral analyses in Python. *J Open Source Softw* 6:2977
- Reagan AJ, Mitchell L, Kiley D, Danforth CM, Dodds PS (2016) The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci* 5:31
- Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-Networks. In: Inui K, Jiang J, Ng V, Wan X (eds), 2019 Conference on empirical methods in natural language processing, Hong Kong, China
- Ricoeur P (1980) Narrative time. *Crit Inq* 7:169–190
- Ryland Williams J et al (2015) Zipf's law holds for phrases, not words. *Sci Rep* 5:12209
- Schmidt BM (2015) Plot archeology: a vector-space model of narrative structure. In: 2015 IEEE international conference on Big Data (Big Data). pp. 1667–1672
- Science Fiction & Fantasy Writers of America (2020) Nebula rules. <https://nebulas.sfwaw.org/about-the-nebulas/nebula-rules/>
- Toubia O, Berger J, Eliashberg J (2021) How quantifying the shape of stories predicts their success. *Proc Natl Acad Sci USA* 118(26)
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems
- Vonnegut K (1999) *Palm Sunday: an autobiographical collage*. Random House Publishing Group
- Vonnegut K (2010) Kurt Vonnegut on the shapes of stories. <https://www.youtube.com/watch?v=oP3c1h8v2ZQ>. Accessed 15 May 2014
- Vonnegut Jr K (2005) *A man without a country*. Seven Stories Press, New York
- Wallace B (2012) Multiple narrative disentanglement: unraveling infinite jest. In: Fosler-Lussier E, Riloff E, Bangalore S (eds) Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada. Association for Computational Linguistics, pp. 1–10
- Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res Methods* 45:1191–1207
- World Science Fiction Society (2022) Hugo award categories. <https://www.thehugoawards.org/hugo-categories/>
- Wu Z, Huang NE (2004) A study of the characteristics of white noise using the empirical mode decomposition method. *Proc R Soc Lond Ser A* 460:1597–1611
- Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal* 01:1–41

## Acknowledgements

The authors are grateful for the computing resources provided by the Vermont Advanced Computing Core and financial support from the Massachusetts Mutual Life Insurance Company. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314. MIF is also grateful to Miguel Fudolig, Joshua Minot, Andy Reagan, and Aritra Banerjee for helpful discussions.

## Competing interests

The authors declare no competing interests.

## Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

## Informed consent

This article does not contain any studies with human participants performed by any of the authors.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-023-01680-4>.

**Correspondence** and requests for materials should be addressed to Mikaela Irene Fudolig.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023