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

Mikaela Fudolig[1], Thayer Alshaabi[2], Kathryn Cramer[1], Christopher M. Danforth[1] and Peter Sheridan Dodds[1]
[1]Vermont Complex Systems Center, University of Vermont, USA
[2]Advanced Bioimaging Center, UC Berkeley, Berkeley, CA, USA

In recent years, quantitative methods have been used to examine changes in word usage in books [1, 2] and have provided support for existing qualitative theories on the progression of narratives [3, 4]. However, these 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.

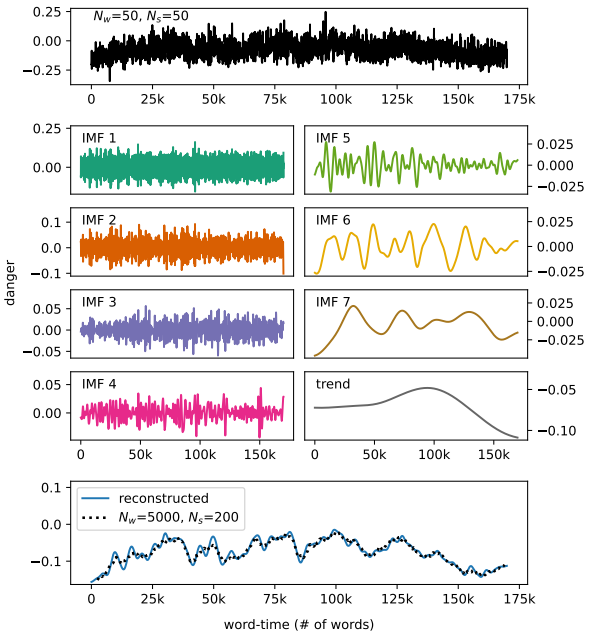

Fig. 1. The top panel shows how danger scores change for "The Iliad" across the course of the text (in number of words). It is decomposed by empirical mode decomposition into a number of oscillatory modes (labeled "IMF") and a non-oscillatory trend. By summing all the modes above IMF 5, the mode at which the original text starts to differ from its shuffled versions, we obtain a denoised version of the time series that is similar to that obtained using larger, overlapping windows.

We use the power-danger framework for essential meaning, or *ousiometrics* [5], an orthogonalization of the valence-arousal-dominance framework derived from semantic differentials, to characterize more than 30,000 books in Project Gutenberg as time series. Each time series is constructed by segmenting the text into non-overlapping windows of length $N_w = 50$ words, with the average danger or power scores for the words in each window corresponding to a point in the time series. The time series thus moves along what we define as "cumulative word-time", measured as the number of words read as one goes through a text. We then perform empirical mode decomposition to decompose the time series into a sum of its constituent oscillatory modes and its general 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.

Authors' note: Our results have been published in [6].

[1] Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

[2] Boyd, R. L., Blackburn, K. G., and Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

[3] Vonnegut, K. (1999). *Palm Sunday: An Autobiographical Collage*. Random House Publishing Group.

[4] Freytag, G. (1900). *Freytag's Technique of the drama: an exposition of dramatic composition and art. An authorized translation from the 6th German ed.* Scott, Foresman, Chicago, 3rd ed. edition. Open Library ID: OL7168981M.

[5] Dodds, P. S., Alshaabi, T., Fudolig, M. I., Zimmerman, J. W., Lovato, J., Beaulieu, S., Minot, J. R., Arnold, M. V., Reagan, A. J., and Danforth, C. M. (2021). Ousiometrics and Telegnomics: 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]*. arXiv: 2110.06847.

[6] Fudolig, M. I., Alshaabi, T., Cramer, K., and Danforth, C. M., and Dodds, P. S., (2021). A decomposition of book structure through ousiometric fluctuations in cumulative word-time *Humanit Soc Sci Commun*, 10:187.