AUGSBURG UNIVERSITY

2211 Riverside Anenue
Minneapolis, MN 55454
612-330-1000
augsburg.edu

Project Final Report: "Signal and Noise and *Pride and Prejudice*: Toward an Information History of Romantic Fiction"

Principal investigator: Dallas Liddle, Augsburg University. **HTRC support**: Eleanor Dickson, Ryan Dubnicek, Nandana Nallapu, Peter Organisciak

Overview and goals

This project pursued two levels of question. On the level of theory, it asked whether the mathematical concept and definition of "information," first developed in 1948 for Information Theory, and foundational ever since for engineers and computer scientists working with technological communications systems, might also help Digital Humanists and literary scholars evaluate the kind of "information" in literary texts. On the level of historical application, it considered whether the rise of the novel in Britain after 1815 could have been related to an increase in the absolute information that innovators such as Jane Austen and Walter Scott were able to code into fictional discourse. In other words, this project asked whether the nineteenth-century novel might have had a heretofore unknown *information history*.

Literary scholarship since the mid-twentieth century has occasionally considered, and rejected, Information Theory as a way of modeling and measuring the content of texts, but the rejection has always been justified on conceptual, not experimental, grounds. In January 2016 the Stanford Literary Lab reported that a corpus drawn from all published British fiction showed markedly different levels of statistical "redundancy" than did a parallel corpus chosen from the literary canon alone. This appeared to be evidence that literary value might have an objectively measurable information profile that could be reflected in the historical record, though the Stanford researchers interpreted their result differently. Because the Stanford work was not peer reviewed, and neither the full composition and results of its corpora nor its methodology was shared, the experiment could not be replicated or its implications futher investigated without creating new experiments and tools.

I proposed with the help of HTRC partly to replicate and partly to pursue the issues raised by the Stanford publication. I hoped to create and study large full-text corpora of British novels, including a reconstruction from HTRC resources of the 250 texts in the proprietary *Nineteenth Century Fiction* collection by Chadwyck-Healey (used by Stanford), and also specialized sub-corpora of the works of Austen, Scott, and individual peer novelists including Edgeworth, Marryat, and Radcliffe. Because my training is in literary studies and my familiary with coding small, I also hoped for assistance to write code to estimate the entropy and redundancy of texts, and to look for patterns of statistically measurable information in fiction.

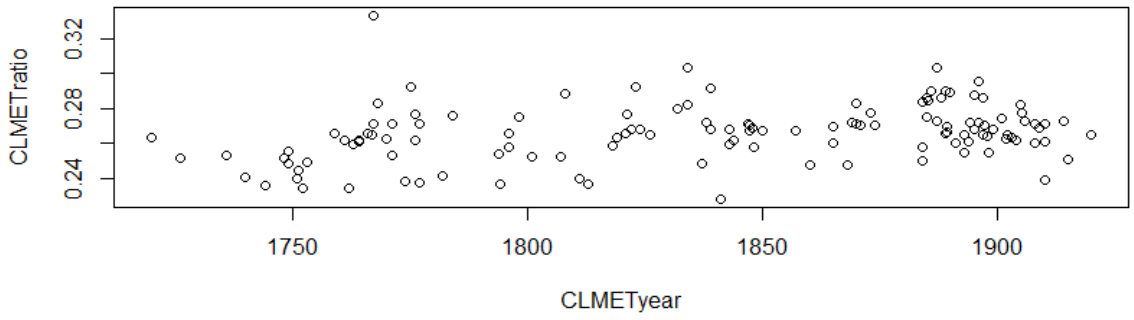
Methods and narrative

When the project was approved in Summer 2016 access was obtained to HTRC text files through an account with the University of Illinois. Eleanor Dickson's assistance was critical in recovering

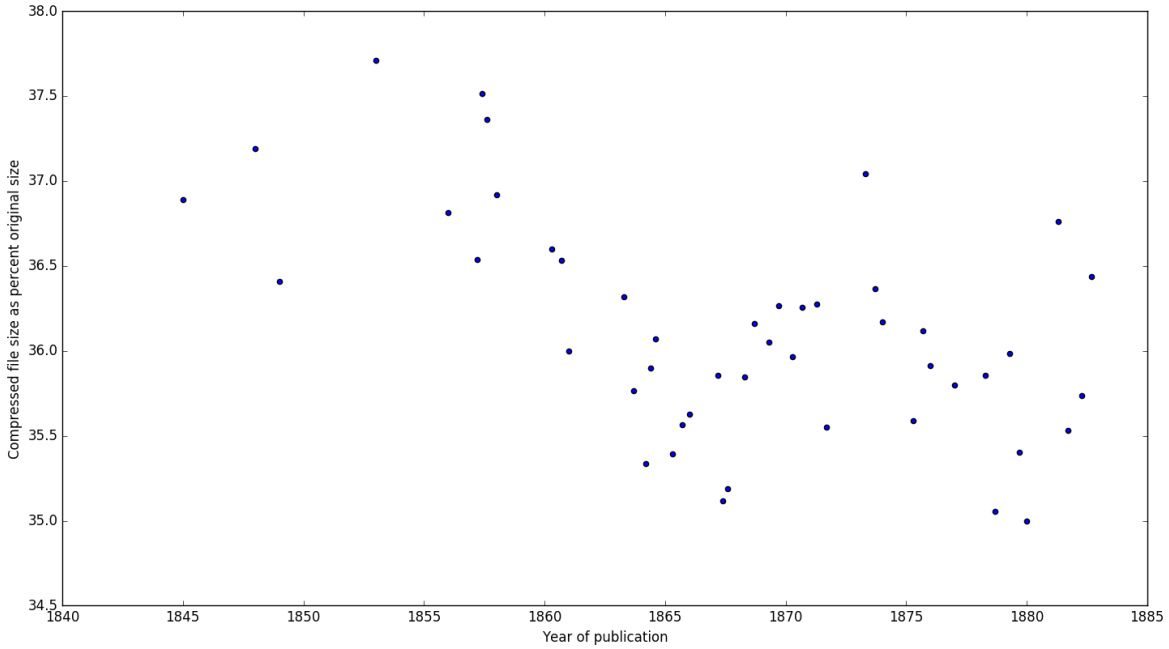
the list of texts that had comprised the Stanford researchers' "canon" and beginning to retrieve versions individually for analysis and incorpora(ation) into searchable resources. Peter Organisciak patiently introduced me to Python by way of Jupyter Notebooks, helped me work through learning resources he had already created, and investigated and shared a range of useful Python resources for textual analysis including TextBlob and Natural Language Toolkit, tools to which I later added a basic understanding of R. Corpora of fiction by Austen, Scott, and Trollope were independently created, and coding tools built to automate creation of specialized sub-corpora including text files containing only the dialogue or only the narrative of in a given fictional work. Tools were also built to compare the rank-frequency (Zipf) traces of multiple novels in a single visualization. As another means of measuring the statistical redundancy in texts, Peter and I settled on the Python version of the file compression tool Zlib, which used Lempel-Ziv and Huffman coding based on Information Theory principles, and he and I developed a simple method in Python of rapidly cycling through large numbers of text files to find their compressibility and plot the change over time of the compressibilility of novels written either by a single author or contained in larger textual traditions.

Outcomes

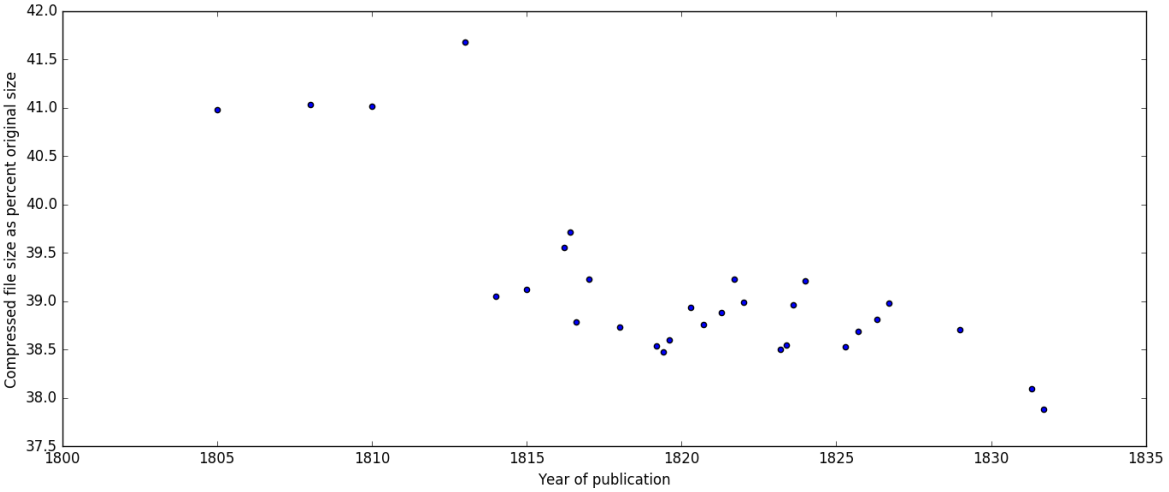
By the end of the grant period the tools and corpora were usable, and I had begun testing the project's major hypothesis on small and large bodies of text. Some months later, five large (150 major texts or more) and multiple small corpora of novels published in the long nineteenth century have been tested, together appearing to show that there was indeed a gradual but replicable and statistically significant increase over the long nineteenth century in the information density of published fiction, as in the example below (compare fiction circa 1750 to that circa 1900) which tests just the fiction in Hendrik De Smet's Corpus of Late Modern English Texts 3.1.



Interestingly, redundancy tests of the individual careers of productive novelists tend to show the opposite pattern over time. Both Anthony Trollope's and Sir Walter Scott's most information-dense fiction comes at the beginning of their careers, with later fiction showing lower levels of information density. In Trollope's case the result is especially interesting, since the dividing line between early high-density and later lower-density works appears to be the historical moment at which Trollope began to contribute to W.M. Thackeray's *Cornhill* magazine and to write fiction using the famously rapid and unrevised method he describes in the *Autobiography* (1883).



Above is the result of the HTRC tool used to study Trollope's information density over his career; below is the same trace for Scott, with his four early long narrative poems included.



The methods developed via the grant have already been used on literally millions of words of comparison corpora including the Hansard archives of debates in the British Parliament.

Scholarly products to date

A monograph in preparation relies heavily on these findings. The following additional work outputs have already been submitted and/or accepted to date:

- A conference presentation (poster), "Textual analysis and the hard problem of interdisciplinary 'information,'" accepted for HASTAC 2017, November 2-3 in Orlando, Florida.
- A manuscript article, "Information Theory and the History of the Novel," under review.