**HIGGINS — Patterns**

<4 images>

**Patterns of Novelty in Literary Data**

Higgins, D. C., Padilla, T. G. and Hintze, A.

In addition to forming a piece of the lasting and living embodiment of the cultural heritage of humanity, literature also constitutes a form of data. The features of this data are precisely what define the ‘literary’ as such. In order to ‘understand the structural continuity of the step from information to literature and back again . . . [and] to grasp the *non*uniqueness of literature in an absolute structural sense’, that is, to specify a difference of *degree* rather than *kind* between literature and other forms of data, it is necessary to isolate and define the features of the literary band of the data spectrum with nuance at a granular level.1

To isolate and define features of literary data, the authors have employed several information-theoretical techniques to analyze literary text and find distinguishing patterns. An algorithm developed to study the information novelty in DNA sequences has been applied to strings of arbitrary text. Previously used to quantify information generated by Twitter users on a daily basis, the algorithm has been adapted here to measure information novelty patterns across fictional texts.

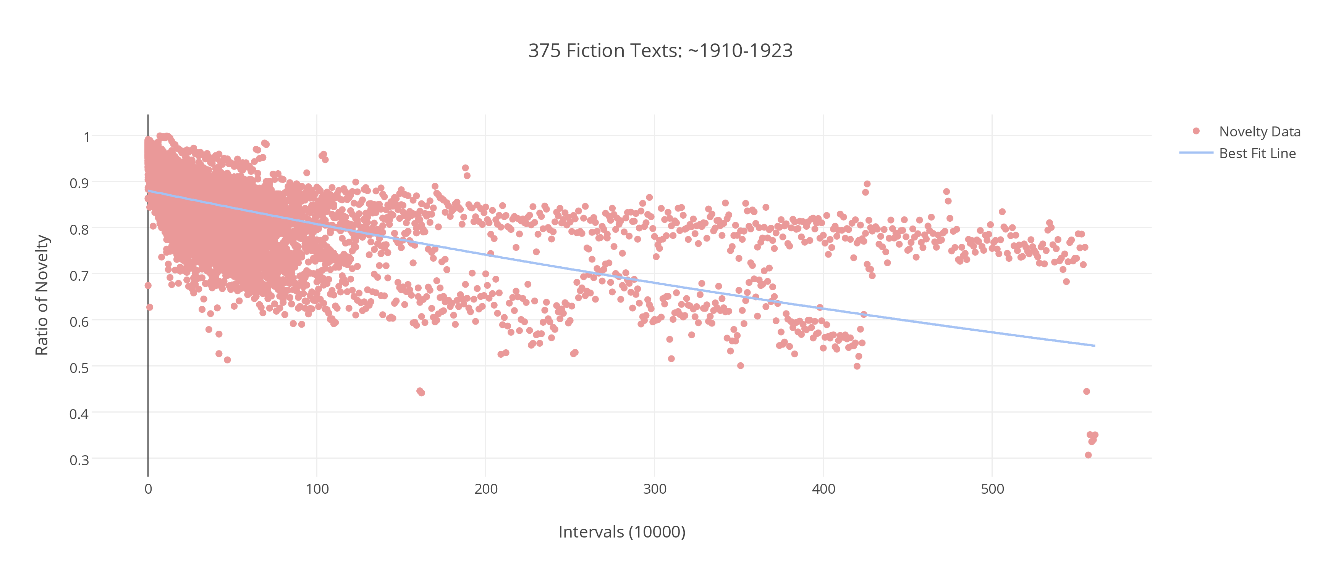


Figure 1. Novelty pattern expressed in 375 texts.

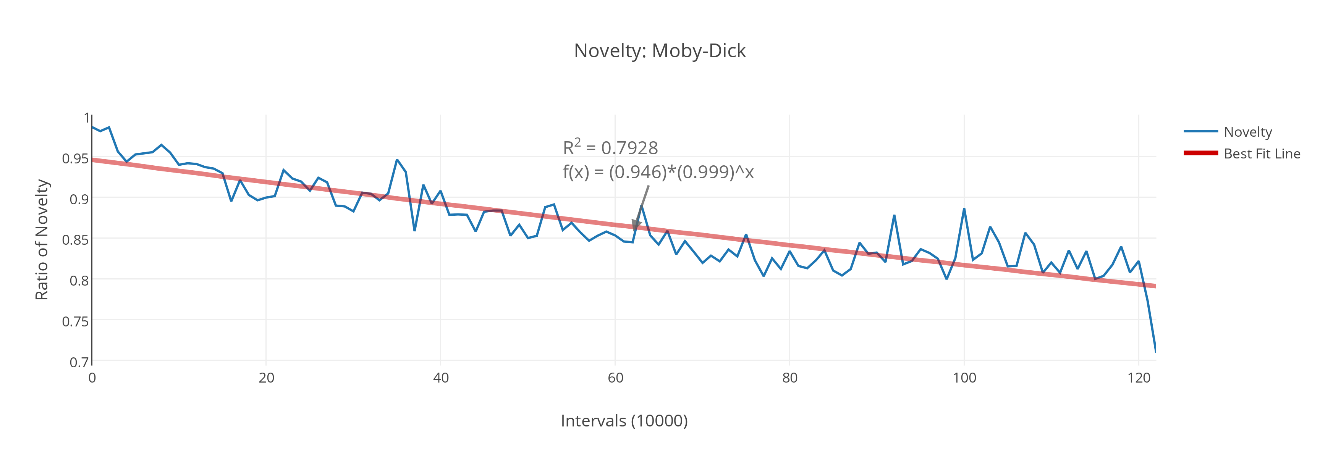


Figure 2. Novelty pattern in *Moby-Dick*.

The graphs above (Figures 1 and 2) measure the proportion of novelty (*y*-axis) over intervals of 10,000 characters (*x*-axis) within each text, moving from beginning to end. Novelty is determined by the percentage of *n*-length character segments that have not previously appeared in a given text. This measurement stands distinct from a measure of lexical diversity wherein a count is given for unique words that occur in a text. The novelty measure accounts for the totality of combinations of characters in a given text rather than counting unique words. In the case of the graphs above, where *n*=5, novelty declines over the duration of texts according to a pattern of exponential decay. Fitting curves to novelty patterns allows us to make quantitative and comparative claims about patterns of information. The r-squared value in Figure 2 indicates that nearly 80% of the novelty data is explained by the exponential function used to describe the curve (in red).

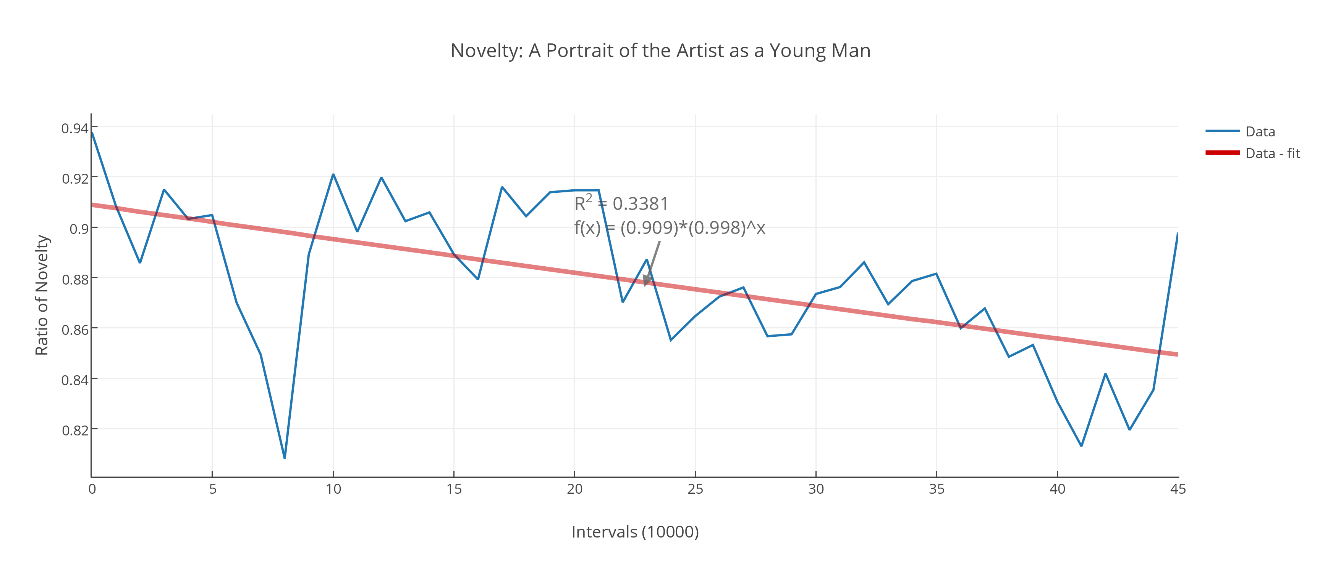


Figure 3. Novelty in *A Portrait of the Artist as a Young Man.*

Yet other texts resist curve-fitting, displaying information patterns that are highly variable and erratic. Figure 3 represents the novelty pattern for Joyce’s *A Portrait of the Artist as a Young Man*, in which the exponential function can only account for approximately 38% of the recorded variation—wild swings in the data that are unexpected, perhaps, in one of Joyce’s less experimental works. (The r-squared value for *Ulysses* was 52%.)

The novelty measure is not only useful when looking at patterns over individual works but as a way of assessing linguistic ingenuity, or fluctuating historical trends in literary authorship, by studying the works over time of a single author, or of many authors across historical epochs. Figure 4 depicts novelty across three novels by Virginia Woolf (in chronological order), in which spikes of novelty are visible at the start of each new work (at approximately the 75 and 175 points along the *x*-axis).

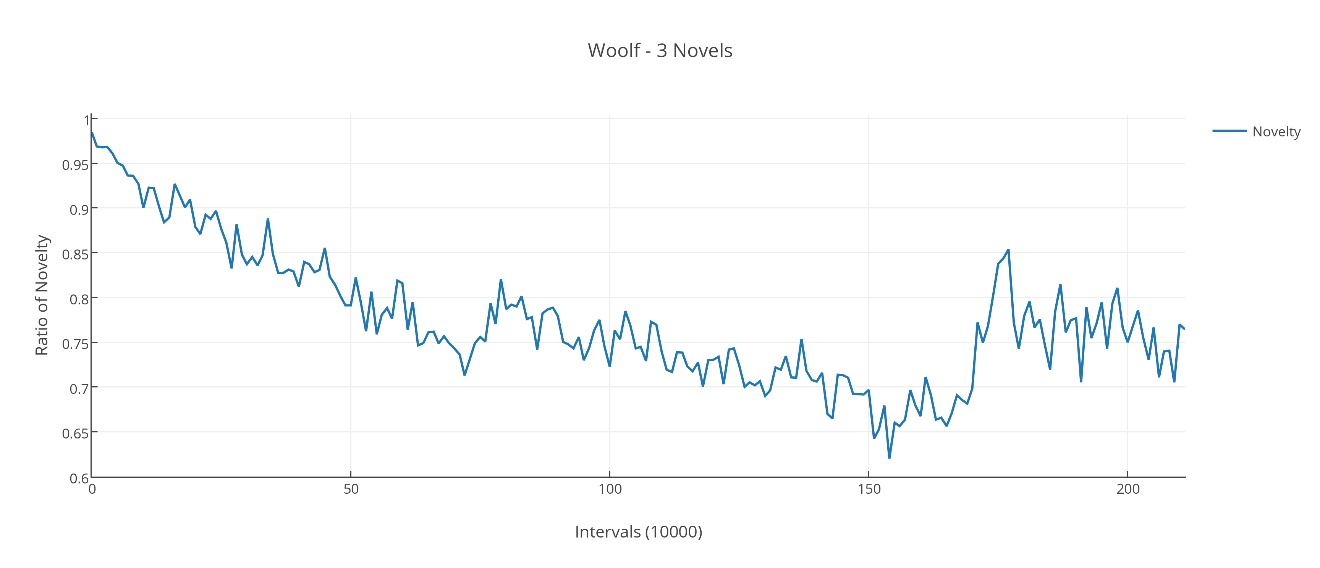


Figure 4. Novelty across three novels by Virginia Woolf.

The significance or not of ‘novelty’ in regard to literary studies is a question for debate that our poster will address. Recent work on patterns of information has shown that the concept of novelty (describing a formation that is new only from a particular perspective) is strongly linked to the concept of innovation (describing one that is new to all perspectives) (Tria et al., 2014). Tying novelty to innovation allows us to go further in building arguments about the role that novelty measurement could play in building an image of the particular form of data known as literature.

Our poster will present visualizations of key findings as we continue to investigate literary data, via an algorithm designed to detect patterns of novelty. The poster would also work well as a live demonstration, during which texts could be fed to the algorithm ‘live’ as the audience circulates and poses questions.

**Note**

1. Terence Turner, quoted in Hayot (2014).

**References**

**Hayot, E.** (2014). What Is Data in Literary Studies?

http://erichayot.org/ephemera/mla-what-is-data-in-literary-studies/.

**Tria, F., et al.** (2014). The Dynamics of Correlated Novelties. *Nature,* http://www.nature.com/srep/2014/140731/srep05890/full/srep05890.html.