Fiction, Non-Fiction, and Family Writing

Investigating the Boundaries of Literary Genre

Eleanor Dumbill

Abstract

The boundaries of literary genres have long been contested. Stylometric investigations of genres — for example, to identify genre through distant reading — is by no means a new area for research. Computational methods are especially useful for large corpora that have not previously been the subject of many enquiries. This might mean the work of a non-canonical author or work that has not been published. Both are true of the primary text used in this paper: a notebook of anecdotes kept by Frances Eleanor Trollope between January 1879 and March 1890. These anecdotes were written in a prose style but were only intended for the consumption of family. While these methods have been used to analyze unpublished works the aim of this research is often to attribute authorship. This paper uses stylometry to compare Trollope's notebook and other family writings with her published works of both fiction and non-fiction.

The boundaries of literary genre have long been contested. The act of naming a genre and defining the aesthetics and forms that differentiate one from another is a nuanced and sometimes highly technical undertaking. The act of grouping texts, however, requires somewhat less human input and might be delegated to a computer program.

This essay uses stylometry (specifically the R package 'stylo') to investigate an unpublished 'notebook of anecdotes' and compares the linguistic properties of this document to those of unpublished letters and published articles written by the same author. It seeks, furthermore, to answer two questions. First, do the three genres have discreet linguistic fingerprints? Second, drawing on the influence of non-literary sources such as philosophy highlighted by Robert Douglas-Fairhurst, can the influence of family writing be seen in published work?

Stylometry is principally associated with authorship attribution, an especially lucrative area of investigation for researchers of Victorian periodicals

Textual Cultures 14.1 (2021): 140–146. DOI 10.14434/tc.v14i1.32854

whose authors were often anonymous. The use of stylometry in genre identification, however, is by no means a new area (Eve 2017, 76–104). In fact, the question of genre can often complicate or enhance authorship attribution. Because stylometric analysis is aimed at "extracting a unique authorial profile", the shifts in this profile that accompany the author's work in different genres can be tracked, without undermining authorship attribution (Eder 2017, 50). In a forthcoming article, Leah Henrickson and I found that works by several members of the Trollope family were successfully attributed to the authors we expected but were grouped according to the genre of the works in a way that we had not expected. Computational methods are especially useful for large corpora that have not previously been the subject of many enquiries. This might mean the work of a non-canonical author or work that has not been published. Both are true of the primary text used in this paper: a notebook of anecdotes kept by Frances Eleanor Trollope between January 1879 and March 1890. These anecdotes were written in a prose style but were only intended for consumption by family.

First, a few words about the practical considerations underpinning my use of stylo. I combined the letters considered in this study into four documents, each comprising all of the letters written within a specific decade. This served two purposes. Many of the letters considered in this paper were short, averaging 343 words. Although stylometric analyses on such short texts are possible, longer test corpora are ideal. By creating longer documents for each decade, I was able to garner more accurate results and question whether the authorship style of Trollope's letters changed over time.

The notebook of anecdotes, which serves as the primary text for this investigation, is an interesting candidate for study. We can be fairly sure that the anecdotes were all written by Frances Eleanor Trollope, using the old-fashioned method of recognizing her handwriting. Though it is possible some of the stories were dictated to her by others, I was reasonably confident from my previous work on Trollope that the contents of this notebook reflected her authorial voice (though this is, of course, by no means a foolproof means of assessing authorship).

There are some oddities introduced by the analysis of private letters. Trollope frequently mixes other languages, most frequently Italian, Latin, French, and Russian, into her writing and uses shortened forms, such as "afftly" for "affectionately". There is also the increased likelihood of unpublished (and therefore unedited) work to include non-standard spellings. These do not preclude successful analysis with stylo. Indeed, in the case of authorship attribution, these non-standard forms can assist in pinpointing an author's unique language use. Further, Eder, Rybicki, and Kestemont developed the tool to be language-independent, and pre-processing can be used to declare a language (Eder, Rybicki, and Kestemont 2016, 107–21). In the case of these texts, I opted for a setting that assessed words such as "don't" as complete units, rather than splitting it into 'do' and 'n't', as is possible with some settings.

The final technical consideration that I should mention is my means of preparing the corpus. The text of the letters and notebook was transcribed from images taken at archives. On occasion, I could not make out a word, and therefore omitted it. I made my best guess at words in Italian and Latin, as I do not personally speak these languages. The text of the published articles and works of fiction was extracted using Optical Character Recognition software and corrected by hand (see Figs. 1 and 2).

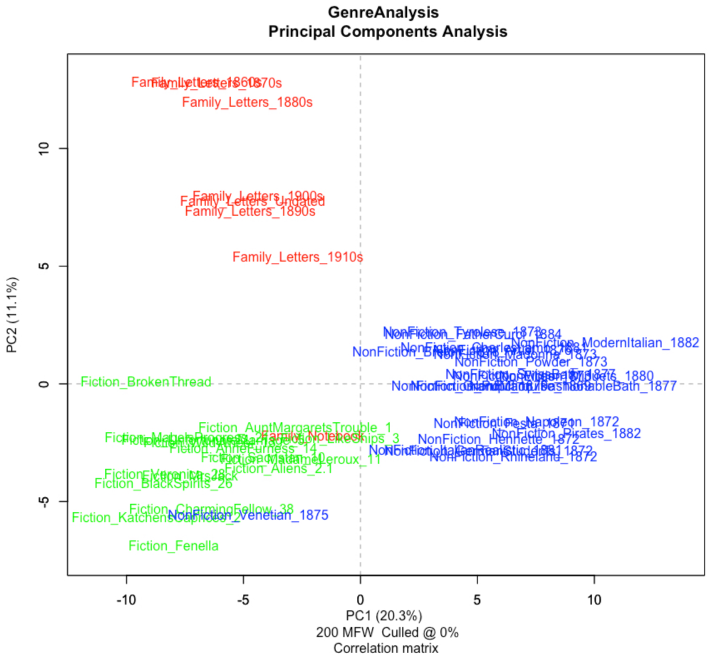

Figure 1. A visualization of the 200 most frequently used words, using principal components analysis. Non-fiction texts are shown in blue, with fiction in green, and family writing in red.

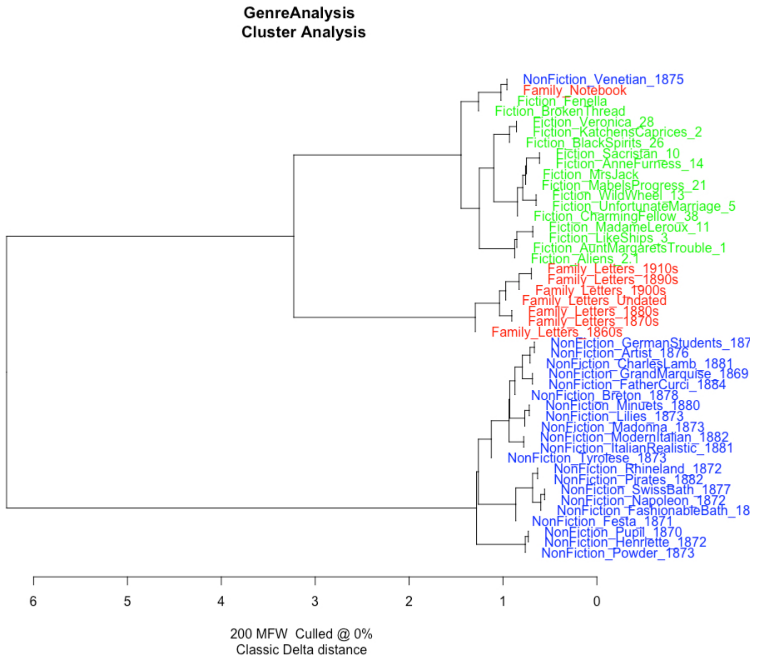

Figure 2. A cluster analysis showing the 200 most frequently used words throughout the corpus.

My corpus comprised 22 non-fiction articles, published in various periodicals, collections of letters covering 6 decades plus a collection of undated letters, 16 works of fiction, and a notebook containing a collection of anecdotes. I have already explained the logic for grouping the letters but should touch on the selection of fiction before moving on. Fourteen of these texts are chapters selected at random from Trollope's novels. Though stylo is able to split longer texts into samples, I elected to manually sample using this method to ensure all the texts I considered were of a similar length and to ease the burden on corpus preparation, already mentioned. Two works of fiction are dissimilar from the others. I included a play, *A Broken Thread* (1903), and Trollope's chapter from the experimental collaborative novel *The Fate of Fenella* (1892). This novel was written using the "exquisite corpse" method whereby authors were not in contact with their collaborators and continued the story without direction from their predecessor.

I expected that the form of these two texts might mean that they were distinguished from the other works of fiction, with *A Broken Thread* differentiated because of dramatic conventions included in the text and *Fate of Fenella* set apart because of the influence of Trollope's collaborators and her attempt to maintain a consistent narrative voice throughout the text.

I limited my assessment of the corpus to two methods. These were the principal components analysis (PCA) and cluster analysis (or dendogram) functions provided by stylo. Both tests provide easily interpreted visual results. With PCA, results are plotted onto two axes (PC1 and PC2) with proximity of a text on both axes denoting similarity in the text's most frequently used words. With the dendogram, texts with similar frequently used words are assigned to the same branch, with the distance between branches denoting similarity — so those that are furthest away or have the most levels between them are most dissimilar.

For both tests, I selected the 200 most frequently used words, without any culling. This is a small number of words, and many studies will choose to use more. Both of these decisions resulted from the relative shortness of my samples. With culling, users can specify the percentage of the corpus in which a word must appear in order for that word to be included in analysis. With 0% culling, which I used, they only need to appear in one text. As noted above, I did not use sampling because of the size of the texts in my corpus.

The results of two of the analyses I performed are shown here.[1] Three clear groupings emerged, showing that, for the most part, stylometric analyses distinguish between the three genres of fiction, non-fiction, and family writing. Other expectations are borne out by these analyses. In both graphs, we can see that *A Broken Thread* and *The Fate of Fenella* are distanced from Trollope's other works of fiction. This confirms the ability of stylo to distinguish upon more specific genre lines than those on the top-level of fiction and non-fiction.

Still when looking at the results of the PCA, *Fenella* and *Broken Thread* are not as distanced according to PC2 (the vertical axis) as the groupings of family letters. These appear in 3 more or less distinct clumps, which appear to be grouped according the date of composition. Those written in the 1860s, 1870s, and 1880s from one group, the undated letters and those from the 1890s and 1900s another, and those written in the 1910s are separate from both. We might conclude from this, for example, that a majority of

[1] Full size images, along with a list of works included in this analysis can be found at https://doi.org/10.17028/rd.lboro.14529279.

the undated letters were written in the 1890s or 1900s. This finding, whose accuracy is somewhat confirmed as they are grouped with these decades in the cluster analysis tree, is still a preliminary suggestion and requires a great deal more investigation. The grouping of the letters by decades also suggests the shifting of Trollope's letter-writing style over time.

My analyses yielded two clear outliers, which were not grouped with their respective genres. The notebook of anecdotes, as expected, sits with Trollope's works of fiction, but so does one of her non-fiction articles, "Venetian Popular Legends", originally published in the *Cornhill Magazine* in July 1875. Looking again at this article, I was quickly able to see why it had been grouped with Trollope's fiction on both the PCA and the dendogram. The article concerns an Italian collection of fairy-tales collected "verbatim" from "old wives [or] gossips" in Castello and Canaregio. This format means that the stories are inaccessible "even [to] those who are very well acquainted with Italian, inasmuch as they are given in unadulterated Venetian dialect" (Trollope 1875, 80). Luckily for the reader, Trollope is proficient enough in Italian to translate the stories, having lived in Italy for much of her adult life. She therefore renders some of the key stories in the collection into English, using her own narrative voice to convey the meaning. This, then, is not a failure of the stylometric analysis but an example of how it can prompt us to reconsider the classifications we have assigned to texts using other means. Having not re-read all of the work included in this study before beginning it (in my own defense, this would defeat the purpose of distant reading), I classified pieces according to their place of publication and a first glance appraisal. This led me to take "Venetian Legends" at face value as a review and therefore a piece of non-fiction. Stylometric analysis encourages me to reconsider this assessment, but, as noted at the beginning of this paper, does not reveal quite *why* this work did not fit with the classification. It is essential that distant and close reading are undertaken together in order to find the answer.

These are preliminary conclusions, and I want to draw this paper to a close by enumerating some of the questions that are not answered here but could be investigated in future research. First, we know that authors are not the only people involved in the preparation of texts for publication. Dickens is known for his hands-on approach to editing and his desire to institute a uniform voice to his journals, a trait that is far from unique to him in the realm of Victorian periodicals. It would be fruitful, then, for a longer iteration of this study to consider whether stylometry can be used to differentiate between Trollope's works published in different periodicals under different editors. This applies equally to fiction as to non-fiction.

Another area of enquiry might consider the authorship of her letters. I gestured earlier to the possibility that the anecdotes included in the notebook may have been dictated by a family member and Trollope frequently wrote letters with her husband, Thomas Adolphus Trollope. Comparing these examples of family-writing with the work attributed solely to both Frances Eleanor and Thomas Adolphus Trollope could provide further insight into the question of family collaboration that I referenced at the beginning of this essay. These are just a few examples of the ways that future work might use similar methods to gain insight into Trollope's writing.

These two tests provided me with several lines for inquiry. Combining the distant reading made possible by computational analysis, close reading, and my existing familiarity with Trollope's life and work I have been able to suggest several conclusions about the texts considered in this study, as well as avenues for future research.

Loughborough University

Works Cited

Douglas-Fairhurt, R. 2002. *Victorian Afterlives: The Shaping of Influence in Nineteenth-Century Literature*. Oxford: Oxford University Press.

Eder, M[aciej], J[an]Rybicki, and M[ike] Kestemont. 2016. "Stylometry with R: a package for computational text analysis". *R Journal* 8.1: 107–21.

Eve, Martin Paul. 2017. "Close Reading with Computers: Genre Signals, Parts of Speech, and David Mitchell's Cloud Atlas". *SubStance* 46.3: 76–104.

Trollope, Frances Eleanor. 1875. "Venetian Popular Legends". *The Cornhill Magazine* 32: 80–90.