Antje Anderson

ENGL 279

Instructor: Jonathan Cheng

29 April 2020

Text-Mining George Eliot's Novels: A Polemic in Favor of Running the Numbers (A Blog Post)

I just ran the numbers on all of George Eliot's novels. In working on my masters' thesis, I had spent a lot of time thinking about her writing on the visual arts at the micro-level—looking at what she had to say about Renaissance art in Florence and about artists like Michelangelo, Raphael, and Fra Angelico. But I was also taking a course on text-mining, where we encountered some databases with tens of thousands of texts and gazillions of words, so I knew that I was looking at a teeny fraction of an already-tiny corpus of seven novels. I began to wonder what forest I was missing by looking at a very few leaves of a rare tree species. I decided to look at the seven novels from a bit of a distance—to use the vantage point of text-mining to check whether some of the claims in traditional Eliot scholarship about the importance of the visual arts in her novels could be quantified in terms of the frequency of certain words about art. At the same time, though, I wanted to keep in mind how few words in her novels I was actually dealing with, numerically speaking.

The main claim that I expected to confirm was that the novel with the most art words and the most words related to Renaissance art would be *Romola*. Eliot's 1863 historical novel about Florence in the 1490s not only mentions many Florentine artists from the 1400s; it references both real and fictional art works from the time and features a real painter, Piero di Cosimo, as a minor character. I wanted to figure out what kind of word frequencies in terms of an "art vocabulary" that would translate into. Beyond *Romola*, though, I couldn't guess whether Middlemarch (1872) or *Daniel Deronda* (1876) would be next in terms of art word clusters: Eliot scholarship places equal emphasis on *Middlemarch*'s iconic scenes in the Vatican Museums and in the painter Naumann's studio, and on the galleries and museums in *Daniel Deronda*, as well as Daniel's own resemblance to a portrait by Titian. And, since these were her last two novels, would a higher frequency of art words in these two mean that "art vocabulary" increased overall in Eliot's writing over time, despite the early spike of art words in Romola in 1863? I knew that a bit of text-mining could show me.

Novel Title	**Year**	**Total number of words**	**Without stop words**	**Art words**	**Art word frequency**
Adam Bede	1859	222,447	89,435	66	0.000297
The Mill on the Floss	1860	213856	87,487	48	0.000224
Silas Marner	1861	74,189	29,777	2	0.000027
Romola	1863	231,391	95,720	134	0.000579
Felix Holt, the Radical	1866	186,502	74,649	50	0.000268
Middlemarch	1872	323,704	129,599	121	0.000374
Daniel Deronda	1876	315,971	125,916	169	0.000535
Total		**1,568,086**	**632,863**	**590**	
Average		**224,012**	**90,409**	**84**	**0.000329**
Average in 19-novel sample		**137,220**	**83,006**	**28**	**0.000208**

Word Count for George Eliot's Novels (art word frequency based on total number of words)

Here is the overall snapshot of Eliot's novels by the numbers. Based on the Project Gutenberg text files of George Eliot's seven novels, which were written between 1859 and 1876, they contain about 1.5 million words, with an average word count of about 224,000. The shortest novel is *Silas Marner* from 1861 (there's always a bit of a squabble about whether to consider it a novel at all), while the longest, *Middlemarch*, contains about 323,000 words. When college students write their classic double-spaced English papers, the ballpark is 300 words per page, so this would translate into a 1,000+ page college paper. Overall, Eliot's novels are typically fairly long even for a "loose baggy monster," as Henry James called the 19th-century novel.[1]

These overall numbers put something about the key passages about the visual arts in Eliot's novels into perspective for me: the number of those prized passages about the arts, and the number

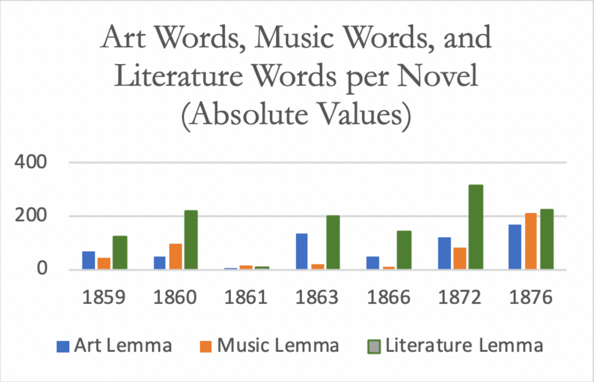
[1] I compared these numbers informally to another small sample of 19 novels written between 1800 and 1900. The average word count for those was only about 137,000 words. The longest novel in this sample was Jane Porter's *The Scottish Chiefs* from 1810, with 306,882 words; the shortest, Maria Edgeworth's *Castle Rackrent*, a little under 30,000. The sample is much too small to be representative, but it affords me a benchmark comparison that I can test on a bigger dataset when I can get my hands on one.

of words throughout her novel that relate to the visual arts specifically, is really, really small. To get some sense of how small exactly, I made a very sophisticated language analysis program used for linguistic machine learning do something very simple.[2] I asked the program to find and count instances of the following ten word stems (so-called lemma) in all her novels: **"paint", "museum", "portrait", "gallery", "artist", "studio", "statue", "sculptor", "sculpture",** and **"picture"**. (I used the lemma so I would get not just "paint," but also "painter," "painted," "painting," and so on.) What did I find? Unsurprisingly, these words were most frequent in *Romola*. So I was able to confirm the claim about *Romola* as the "artiest" Eliot novel. The next novel in line in terms of word frequency was actually *Daniel Deronda*, followed by *Middlemarch*—mystery solved. And just to get math-y for a minute, I ran a linear regression on the frequencies per novel, which showed that the use of these words overall did slightly increase over time, the slope flattened significantly by *Romola*.[3]

Because I really didn't think this information was useful in isolation, I ran similar numbers on a word cluster for music (for which the frequencies are somewhat similar to those for art—and of course there are Eliot scholars writing about her interest in music). I created a word cluster for literature as well, which unsurprisingly showed that words having to do with books and writers are more frequent than either those about art or music. I've included some visuals that show the word frequencies for all three. But the comparison across word clusters is always a little iffy without more computational linguistics: what is the equivalent of "painter" for music? For literature? How is a concert related to a museum or a library? All these are questions that can at least be partially answered by way of calculating the proximity of words, but I haven't done this kind of work here.

[2] For people who want to know the tech specs, I used a Python package called SpaCy with the linguistics "toolkit" NLTK to parse the texts of these novels, and then built and ran my queries and my analysis on the resulting dataframe with Pandas, another Python library.

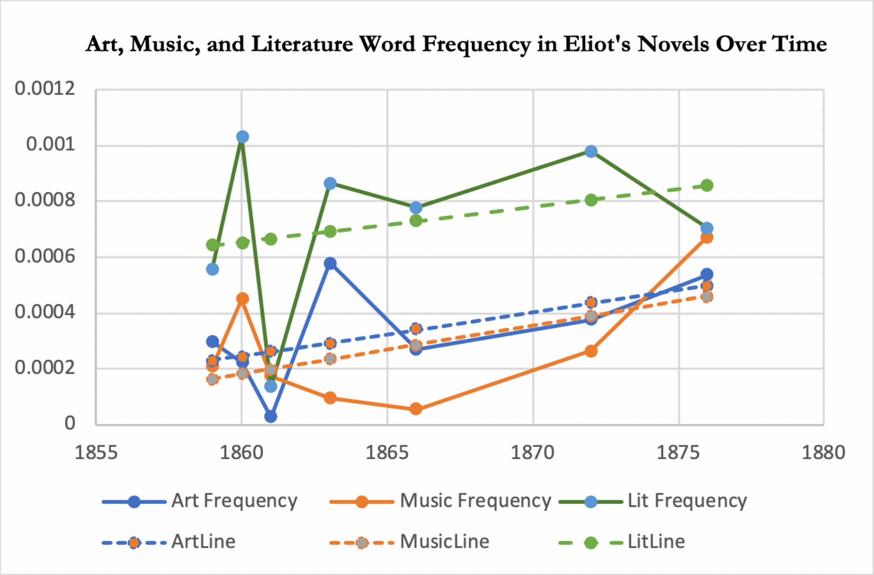
[3] The linear regression was run with the help of another Python library designed for machine learning, SciKitLearn.



Art Lemma: "paint" "museum" "portrait" "gallery" "artist" "studio", "statue" "sculptor" "sculpture" "picture"

Music Lemma: "music" "sing" "piano" "opera" "concert" [other instruments were omitted because they were rarely used]

Literature Lemma: "author" "book" "novel" "library" "manuscript" "write" "poem"



But here is what I want to stress: Even the "most frequent" use of these words in *Romola* came down to a total of only 134 of these words out of over 230,000—not quite .06%, or 6 words per 10,000. For *Daniel Deronda*, the frequency is not very different, with 5 words per 10,000; *Middlemarch* surprisingly lags behind a little at not quite 4 art words per 10,000. With these as the top three, the average across all seven novels came down to 0.03%, or 3 per 10,000 words.[4] Arguably, these word clusters are a very crude measure. Art words do not show up in isolation, and I should properly count other words connected to them that generate larger "art statements" at the sentence

[4] When I asked the program take out what are called the "stop words"—the words that do not carry "content" in themselves, like "the" or "of" etc.—art words rose in frequency, of course, but in *Romola* they still only amount a little more than 0.1%, or 13 words per 10,000.

and even paragraph level. But then again, these words will sometimes be clustered in the very same sentence, while in other contexts, a word like "picture" might not relate to art at all, if used metaphorically (I did not even include the word "art" itself because it is used in so many different nonvisual contexts). But crude measure or not, what remains is that the proportion of these words is very small in all of Eliot's novels. Based on the occurrence of the words from this vocabulary cluster, it is hard to argue (as scholars invested in Eliot's use of the visual arts do) that Eliot's novels particularly emphasize art—and, perhaps just as important, it is hard to argue that they are very different from other novels when it comes to art. Admittedly, Eliot does use more art words than the writers from the sample of 19 novels across the century I mentioned before: in those, there are only 2 art words per 10,000 words on average.[5] But how can Eliot scholars (myself included) justify making claims about the importance of Titian's portraits for *Daniel Deronda* or of Fra Angelico's frescoes in Florence for *Romola* when these references to this art are so few and far between?[6]

Running these numbers on Eliot's novels paradoxically would strike both text-miners *and* traditional scholars as silly. In terms of text-mining, a of seven novels is puny, as is one of 27 novels (Eliot's plus my "control group"). For context: HathiTrust is a digital library of over 8.3 million book titles, with over 600,000 English-language books that date from between 1800 and 1900 (although this number is a little iffy, because while it doesn't count periodicals, it does count duplicates, and there are many). And even if I go by the old ballpark figure of 10,000 novels published in English between 1800 and 1900, probably a low estimate, seven novels is still too small

[5] This is true even though one novel, Catharine Sedgwick's *Clarence* (1830), features about 6 art words in 10,000, just like *Romola*. But I also kicked one novel out of what was originally a sample of 20 novels, because it was such an outlier. That novel is Oscar Wilde"s *Portrait of Dorian Gray*, where the frequency of the art words is larger by a magnitude than the average for the rest, at 23 words per 10,000, and still almost 4 times that of *Romola*. Given the content of the novel, this makes complete sense: *Dorian Gray* turns out to be a handy benchmark for what word frequency might look like in a novel that is *literally* about a work of art, as well as about painters, their studios, and their patrons (among other things).

[6] References to particular artists are even less common in Eliot, at least beyond *Romola*. Titian, Raphael, and Leonardo are the only artists from the Renaissance to be mentioned in her novels (in a total of nine references over three novels!), even though Eliot loved them and wrote frequently about seeing their artwork in her travel journals.

a sample to really bother with—less than 0.1%. So taking the view from 10,000 feet, analyzing these seven novels wasn't going to yield any generalizable findings about "the 19th-century novel"—which is precisely what a good text-miner would want to do.

On the other end, from the close-up vantage point of Eliot experts, creating word clusters and trying to quantify Eliot's vocabulary does not add anything to our discussion of her novels that analysis "by hand" wouldn't address with more nuance. Even traditional scholars who are not categorically opposed to the computational analysis of literature would point out that nothing can replace READING the novels and analyzing them with close attention to detail (as I have done for years). And with only seven novels, that is very doable. What would getting a bunch of statistics about word use and vocabulary clusters add to that? Certainly nothing more than looking for the number of specific instances of word use relating to art in the ancient KWIC concordance, a handy tool available on the internet since the 1990s. The apparatus of text-mining seems rather oversize and unwieldy for the purpose of looking at Eliot's stack of seven.

I can accept all that. But here are the two reasons why I checked out the numbers on Eliot's use of art terminology in her novels anyway. First of all, I do agree with Ted Underwood, eminent scholar among the text-miners (even as I make this point at the small scale of 7 novels by one author, as opposed to up churning 100,000+ novels through his programs as he sometimes does): If we never put long-standing generalized truths about "the novel" that rest solely on a series of close readings of canonical novels to the test, we run the risk of missing the big picture. Given Eliot's towering canonical status, major generalizations regarding the role of visual culture in the Victorian novel are at stake. In other words, there may be nothing wrong with our close readings (I stand by my readings of famous art passages in Eliot), but they may not have any significance and cannot be generalized. The miniscule role that "art words" are playing even in her own novels are a good reminder of that. Even as I can say with confidence on the basis of biographical evidence that Eliot was fascinated by visual art and by specific artists, the small scale at which this interest translates into her novels (even the most "artsy" one) might suggest that we are exaggerating its importance.

Secondly, and perhaps more importantly, my defense of probing any given writers' work with some text-mining tools is much more fundamental: Why the hell *not* text-mine, as a standard practice? I frankly do not understand why we would not include, in any analysis hinging on the WORDS of a novel for whatever purpose, a snapshot of word frequencies, word relationships, sentence length, and other basic statistics on style and usage, made incredibly easy for us to generate with the tools of computational linguistics. Text-mining should simply be part of the standard methodological toolbox. If I write about a poem, even if my focus is not specifically the poem's meter, rhyme scheme, or overall form, I will still do due diligence and complete a scansion, analyze the rhyme scheme, and track assonances, consonances and other sound features of the poem. I do this even when only an offhand sentence about these formal features ends up in my reading of the poem. Why shouldn't I do the same for a given novel's linguistic structure, ideally in comparison with a large sample of other novels from the same time period?

My basic and tentative probing of Eliot's seven novels may neither satisfy the DH scholars who do this kind of work (like Ted Underwood or Matt Jockers), nor the traditionalist scholars who think that all this tinkering with frequencies and averages and regression lines is a waste of time. But I think it might constitute the kind of common "middle ground" that would allow us to maneuver the conceptual space between looking from a distance at the gigantic forest of books and inspecting the ribs on an individual leaf of one specimen of a subspecies under microscope. And that sort of middle ground is where I'd like to dwell, as a reader and miner of books.

Bibliography

Anderson, Antje. "Gendering Art History in the Victorian Age: Anna Jameson, Elizabeth Eastlake, and George Eliot in Florence." MA Thesis, University of Nebraska-Lincoln, May 2020. https://digitalcommons.unl.edu/artstudents/147.

Capuano, Pete. "Racial Science and the Kabbalah in Eliot's *Daniel Deronda*." *Changing Hands: Industry, Evolution, and the Reconfiguration of the Victorian Body*, 152-182. Ann Arbor: University of Michigan Press, 2015.

Eliot, George. *The Journals of George Eliot*. Edited by Margaret Harris and Judith Johnston. Cambridge: Cambridge University Press, 1998.

Eliot, George. *Romola*. Edited by Andrew Sanders. London: Penguin, 1980. [Orig. pub. 1863.]

Jockers, Matt. *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, and Springfield, Illinois: University of Illinois Press, 2013.

Manning, Christopher D. and Hinrich Schütze. *Statistical Natural Language Processing*. Cambridge, Mass., and London: MIT Press, 1999.

Matsuoka, Mitsuharu, "George Eliot." *The Victorian Literary Studies Archive: Hyper-Concordance*. http://victorian-studies.net/concordance/eliot/.

Ormond, Leonee. "Angels and Archangels: *Romola* and the Paintings of Florence." In Caroline Levine and Mark W. Turner, editors. *From Author to Text: Re-Reading George Eliot's* Romola, 180-190. Aldershot: Ashgate, 1998.

Ronald, Ann. "George Eliot's Florentine Museum." *Papers on Language and Literature*, 13 (1977): 260-69.

"Statistics and Visualizations." HathiTrust. https://www.hathitrust.org/statistics_visualizations.

Underwood, Ted. "Preface: The Curve of the Literary Horizon." ix-xxii. *Distant Horizons*. Chicago: University of Chicago Press, 2019.

Underwood, Ted. "We Don't Already Understand the Broad Outlines of Literary History." Blog Post. February 8, 2013. https://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/.

Witemeyer, Hugh. *George Eliot and the Visual Arts*. New Haven and London: Yale University Press, 1979.