Chapter 3

Computational Analysis

3.1 Rigorous, Retrievable, and Replicable

Throughout the computational analysis of my corpus, I follow Simpson's criteria for stylistic analysis, as outlined in *Stylistics: A Resource Book for Students.* Simpson states that stylistic analysis must be *rigorous*, in that it must be "based on an explicit framework of analysis", *retrievable,* meaning that it is "organised through explicit terms and criteria", and *replicable*, or in other words, "sufficiently transparent as to allow other stylisticians to verify them, either by testing them on the same text or by applying them beyond that text."

In keeping with these criteria, and in order to ease future replicability and maintain transparency, in this chapter, I outline the processes of corpus building, text-processing, and analysis that were followed while conducting the computational exploration of my corpus. Additionally, I recognize that computational text analysis involves a variety of choices made on the part of the researcher – choices that ultimately affect the nature of the results. Therefore, I aim not only to to provide a detailed account, but also a justification of the choices that I made when preparing and analyzing my texts.

3.2 Corpus Building

Before outlining the methods by which texts were analyzed, I must first describe the process of selecting texts for inclusion in the corpus. My goal was to create a corpus that would help me answer the questions that I cam seeking to answer, namely: *How does the rhetoric contained within the Thomsonian medical movement differ from that of the orthodox medical community?* In order to do this, it was necessary to build a corpus that

contained a curated variety of medical texts from the 19th century, keeping in mind the historical context in which the texts were created. In curating this set of texts, I aimed to emulate Katherine Bode's concept of a scholarly edition, which she defines as "a model of literary works that were published, circulated,and readand thereby accrued meaningin a specific historical context, constructed with reference to the history of transmission by which documentary evidence of those works is constituted." (Bode, Conjectures)

Because the the unique legal history of the Thomsonian movement, I chose to incorporate not only medical manuals associated with the movement, but also a variety of periodicals and lectures that followed the Thomsonian system. Additionally, because I aim to compare Thomsonian rhetoric to that of more orthodox medical communities, I also chose to include a variety of journals that were associated with state-affiliated medical societies. In this way, my corpus brings together texts that were published and circulated within the American medical community of the 19th century.

The state-affiliated medical journals were sourced from the State Medical Journals collection, which is housed within the Medical Heritage Library. These texts were accessed using the Internet Archives API. The State Medical Journals collection contains texts spanning from 1845 to 2017, but for the purpose of this study, I extracted only texts that were published between 1845 and 1865. Because this collection does not contain texts from the early 19th century, I also hand selected pre-1845, state-affiliated journals from the Historical Medical Journals collection, housed within the Medical Heritage Library. Because there is no pre-existing, curated collection of Thomsonian texts, in order to retrieve this portion of the corpus, I extracted all items in the Internet Archive which contained the term "Thomsonian" in either their title or description, and were published between the years 1824 and 1863. I then hand-sorted through the results and discarded any irrelevant texts ie: texts that used the term Thomsonian, but were not affiliated with Samuel Thomson's medical system.

In addition to compiling texts in a machine readable format (.txt), I also acquired metadata for all texts. When available, the following metadata fields were acquired: *Collection, Contributor, Creator, Date, Description, Format, Genre, Identifier, item size, Mediatype, Public date, Publisher, Rights, Source, Subject, Title, Type, Volume,* and *Year*. For a table containing the texts within the corpus, as well as all corresponding metadata, see the association GitHub repository. The corpus contains a total of 59 texts.

After texts and their corresponding metadata were acquired, the corpus was broken down into four groups, with each group containing texts with distinct differences in their audience and purpose. Table 1 contains a breakdown of each group and its contents:

Class (abbreviation)	Count	Description
ThomsonianManuals (tm)	18	Texts that serve to describe the Thomsonian system and its applications. Generally aimed at the laymen for practical use.
ThomsonianPeriodicals (tp)	7	Serial publications, aimed at both presenting case studies and updating practitioners on the system.
ThomsonianCommentary (tc)	16	Testimonies, lectures, reports, etc discussing the merits of the system in comparison to orthodox practices.
MedicalJournals (j)	18	State affiliated serial publications, containing cases aimed at physicians.

3.3 Text Processing and Cleaning

Before conducting analyses, texts in the corpus were cleaned and processed using the tm and NLP packages in R. The corpus was transformed to lowercase, and all numbers and punctuation were removed. In keeping with standard practices of corpus preparation, English stopwords were removed. A full list of stopwords can be found in the documentation for the tm package. Additionally, the corpus was stemmed so as to remove variance created by different forms of the same word. (For example, prior to stemming, the computer will recognize "doctor" and "doctors" as two different and unrelated tokens. This is not advantageous in analysis.) While the removal of numbers and punctuation changes the ultimate composition of the corpus, I chose to remove them due to the nature of the questions that I am asking. The script that was used to clean the texts can be found in the following GitHub repository:

`https://github.com/abigailella/dh-thesis`

3.4 Principal Components Analysis

Principle components analysis (PCA) is a means of summarizing multidimensional data while maintaining the relationship between variables. It is sometimes called "eigenvector analysis" or "latent vector analysis". [1] PCA is used frequently in Computational Stylictics, a field of study which "aims to find patterns in language that are linked to the processes of writing and reading, and thus to 'style' in the wider sense" [2]

3.4.1 A Brief Mathematical Explanation

Before exploring the findings of a PCA run on my corpus, I find it necessary to give a brief explanation of how PCA works within the context of computational text analysis. PCA uses "the frequencies of occurrence of p function words in n text blocks as input for the analysis."[3] In other words, the analysis must begin with a matrix of the frequencies for each word in a corpus, across each word in that corpus. For clarity, I've provided an example of what such a matrix might look like, below.

	life	**nature**	**cell**	**right**
text_1	2	3	4	0
text_2	2	5	0	4
text_3	1	3	1	3
text_4	4	2	1	2
text_5	1	6	0	2

This matrix shows us that the word "nature" occurs in text_1 three times, compared to text_2, where it appears five times. A matrix like the one shown above is the starting place of PCA. However, imagine that this matrix was expanded to include every word that happens to occur within a corpus of dozens, hundreds, or thousands of novel-length texts. Such a matrix would be so large that we'd be unable to glean any useful information from it. In layman's terms, PCA helps to reduce the dimensionality (ie: the *size*) of the matrix so that it may be more easily summarized and visualized, while not erasing meaningful information contained within the matrix.

Starting from the matrix of raw word frequencies, we must first calculate the *mean corrected* and *standardized* frequencies. The *mean corrected* frequencies are calculated by taking the frequency of a given word in a single text and subtracting the average frequency of that word across the entirety of the corpus.

This process is repeated for each text in the corpus. The *standardized* frequency, also calculated for each text, takes the *mean corrected* frequency and divides it by the standard deviation of the word across the corpus. The standard deviation is simply a measure of "spread" in the data. In the example matrix provided above, "life" has a spread from one to four across the five text, whereas the word "cell" has a spread from zero to four. "Cell" will therefore have a higher standard deviation than "life". For reference, the standard deviation for any given word across the corpus is given by the following formula, where x represents the *mean corrected* frequency and n represents the total number of texts in the corpus:

$$s = \sqrt{\frac{\Sigma(x)^2}{n}} \tag{3.1}$$

In layman's terms, this gives us a sense of the "spread" of the data. We can then calculate the variance for each word by taking the square of the standard deviation of the word's frequency across the corpus. In other words, s^2.

However, it is not particularly helpful to look only at each word's individual variance without also examining how this variance compares to the variance of each of the other words contained in the corpus. In order to understand how each word (or rather, *column of word frequencies*) in the matrix contributes to the overall variance within the corpus, we must look at measures of covariance. The covariance between any two columns in the matrix is given by the following formula, where x_1 is the mean corrected frequency of the first column of words and x_2 is the mean corrected frequency of the second column of words:

$$s = \frac{\Sigma(x_1 * x_2)}{n} \tag{3.2}$$

In a corpus, however, there are far more columns than merely x_1 and x_2, as there are far more than two words contained within any given corpus. We must calculate the covariance for all *all* possible word pairs within the corpus. Luckily, covariance across a corpus can be plotted in a matrix.

Even so, we are left we yet another large matrix. Principle Components Analysis attempts to reduce the dimensionality (or *size*) of our covariance matrix on the basis of which components (in this case *words*) contribute the most variance to the data set. The combination of words contributing most to the variance is called the first principle component, or *PC1*, and the combination of words contributing the second most to the variance is called the second principle component, or *PC2*. Components of the corpus that do not make a significant contribution to variance are discarded in the process of PCA, thereby reducing the dimensionality of the matrix.

When we plot the results of our principle components analysis, we end up with a graph situation around the axis of the first two principle components. On a PCA graph, each point represents a text. The farther the distance between the points, the more different they are based on their principle components.