Stylometry:

The Transformation of Words into Numbers

By Mya Khela

Preface

In 1818, Frankenstein or otherwise known as the Modern Prometheus was first published. This chilling tale first created for a ghost story competition between friends has become one of the most prevalent stories in literature, founding the genre of science fiction!

While today, Mary Shelley is recognised as the true author of Frakenstein, over the years there has been some speculation as to whether this is true. One concern for this is the fact that the novel was published anonymously, only giving Mary credit as the author in the 1821 publication. Although it was quite common for 18-19th century writers to remain anonymous, many still thought her husband, Percy Shelley was the writer.

Belief for this escalated when it was found that Percy had in fact written the preface of the novel along with transcriptions found of the original manuscripts showing parts written in Percy's handwriting!

But with all these contrasting opinions, how could we show that Mary was the true author?

What if I told you the answer is down to maths...

Chapter 1

Stylometry is the quantitative analysis of literature, a science first introduced by British mathematician and logician Augustus de Morgan in 1851 who suggested that authorship could be determined by ananlysing the average number of letters written by an author. However, the actual term 'stylometry' was coined by polish philosopher Wincenty Lutosławski in the 1890s during his research in organising Plato's dialogues.

Stylometry relies on the idea that every author's writing has specific characteristics that remain fairly constant throughout all of their works. These characteristics range from average sentence length to the arrangement of words to even the frequency at which words occur.

Over the years, this science has flourished, helping to solve many authorship disputes that the usual literary analysis couldn't solve. For example in 1964, Mosteller and Wallace were able to use stylometric analysis to conclude that James Madison wrote the 12 disputed federalist papers that Alexander Hamilton had claimed to have written himself. Further studies include determining whether Shakespeare actually wrote his own plays and whether Lenon or McCartney wrote the Beatles song 'In My Life', along with it also being used by forensic analysts to solve crimes!

So how might stylometry be able to help us in the question of Frankenstein's authorship?

Well by using one of the most well known stylometric techniques, Burrow's Delta Method!

Chapter 1.1

The delta method, introduced by John Burrows in 2002 is used to find the distance between an unknown text and a set of candidate texts, written by those most likely to be the author by analysing the most frequent words in the texts.

The method relies on a writer's use of function words. Function words, are words that have very little meaning on their own such as 'the', or 'and'.

But what makes these meaningless words so interesting and important to stylometric analysis?

Function words can form a reliable data set for each author due to the fact that all writers use them unconsciously, which is a trait that makes it very hard for others to copy! This creates something like an 'authorial fingerprint', which sets each writer apart. What also makes the use of function words great is that they are used so much throughout texts that they provide a large dataset for us to use!

Now I'll run through how the delta method actually works, using an example to show how it relates to the Frankenstein authorship problem...

According to the method, we first have to assemble a corpus, which is just a collection of the texts we will use in the analysis.

For our corpus, I've decided to use Valperga by Mary Shelley, St Irvyne by Percy Shelley and Frankenstein which will be the 'unknown text'.

Then I found the 30 most frequent words in the corpus and recorded the frequencies of the words in each text as shown in the table below:

	FREQUENCY			
WORD	CORPUS	VALPERGA	FRANKENSTEIN	ST IRVYNE
the	9981	3724	4186	2071
of	6279	2371	2640	1268
and	6133	2294	2972	867
to	4424	1440	2092	892
I	3699	485	2850	364
a	2904	1062	1386	456
in	2374	791	1129	454
was	2090	701	1021	368
my	2191	288	1776	127
that	1991	677	1017	297
with	1513	591	667	255
had	1347	413	685	249
as	1111	397	528	186
by	1067	415	459	193
it	1114	287	545	282
for	1081	369	497	215
which	1306	365	558	383
on	1003	320	460	223
but	1189	319	686	184
not	1031	320	510	201
me	1107	125	867	115
from	827	259	384	184
you	1064	307	574	183
at	737	249	317	171
were	759	302	316	141
this	706	231	402	73
be	704	236	358	110
we	282	73	171	38
is	663	196	305	162
their	614	344	184	86
TOTAL WORDS IN TEXT	160517	54006	74919	31592

Chapter 2

For the next step, we need to calculate the mean frequency of each word across the corpus of known authors (St Irvyne and Valperga).

The reason why we don't include the unknown text in the calculation is because the mean will be used to help calculate the distance between the known authors and the unknown text later on. If we use the unknown text in the mean, then the results could be skewed and the authors would appear to be closer to the unknown text than they are.

The mean, $\bar{x}$ is just the average value in a dataset, found by dividing the sum of the values in the set by total number of values, which can be expressed as:

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

Where n = the number of values and x_i = represents an individual value in the dataset.

To calculate the mean frequency, we first have to calculate the relative frequency of each word in each text.

The relative frequency, is the ratio of the number of times an event occurs to the total number of observations.

So in our case:

$$\text{Relative Frequency} = \frac{Number\ of\ times\ word\ appears\ in\ text}{Total\ number\ of\ words\ in\ text}$$

For example, for the word 'the', its relative frequency in Valperga is:

$$\frac{3724}{54006} \sim 0.0690 \sim 6.90\%$$

and in St Irvyne:

$$\frac{2071}{31592} \sim 0.0656 \sim 6.56\%$$

If we repeat this for the other words we get...

	RELATIVE FREQUENCY		
WORD	VALPERGA	FRAKENSTEIN	ST IRVYNE
the	6.90%	5.59%	6.56%
of	4.39%	3.52%	4.01%
and	4.25%	3.97%	2.74%
to	2.67%	2.79%	2.82%
i	0.90%	3.80%	1.15%
a	1.97%	1.85%	1.44%
in	1.46%	1.51%	1.44%
was	1.30%	1.36%	1.16%
my	0.53%	2.37%	0.40%
that	1.25%	1.36%	0.94%
with	1.09%	0.89%	0.81%
had	0.76%	0.91%	0.79%
as	0.74%	0.70%	0.59%
by	0.77%	0.61%	0.61%
it	0.53%	0.73%	0.89%
for	0.68%	0.66%	0.68%
which	0.68%	0.74%	1.21%
on	0.59%	0.61%	0.71%
but	0.59%	0.92%	0.58%
not	0.59%	0.68%	0.64%
me	0.23%	1.16%	0.36%
from	0.48%	0.51%	0.58%
you	0.57%	0.77%	0.58%
at	0.46%	0.42%	0.54%
were	0.56%	0.42%	0.45%
this	0.43%	0.54%	0.23%
be	0.44%	0.48%	0.35%
we	0.14%	0.23%	0.12%
is	0.36%	0.41%	0.51%
their	0.64%	0.25%	0.27%

NOTE: Although the relative frequency for the words in Frankenstein won't be used to calculate the mean, I have calculated them anyway as they will be needed later on in the analysis.

Now that all the relative frequencies have been calculated, we can work out the mean frequencies:

For 'the', using the equation for the mean, we have to sum the relative frequencies and then divide it by how many values there are which is 2, giving us...

$$\bar{x} = \frac{0.0690 + 0.0656}{2} \sim 0.0673$$

Repeat this with the other words and we get:

	RELATIVE FREQUENCY			
WORD	VALPERGA	FRAKENSTEIN	ST IRVYNE	MEAN FREQUENCY ACROSS CORPUS
the	6.90%	5.59%	6.56%	0.0673
of	4.39%	3.52%	4.01%	0.0420
and	4.25%	3.97%	2.74%	0.0350
to	2.67%	2.79%	2.82%	0.0274
i	0.90%	3.80%	1.15%	0.0103
a	1.97%	1.85%	1.44%	0.0170
in	1.46%	1.51%	1.44%	0.0145
was	1.30%	1.36%	1.16%	0.0123
my	0.53%	2.37%	0.40%	0.0047
that	1.25%	1.36%	0.94%	0.0110
with	1.09%	0.89%	0.81%	0.0095
had	0.76%	0.91%	0.79%	0.0078
as	0.74%	0.70%	0.59%	0.0066
by	0.77%	0.61%	0.61%	0.0069
it	0.53%	0.73%	0.89%	0.0071
for	0.68%	0.66%	0.68%	0.0068
which	0.68%	0.74%	1.21%	0.0094
on	0.59%	0.61%	0.71%	0.0065
but	0.59%	0.92%	0.58%	0.0059
not	0.59%	0.68%	0.64%	0.0061
me	0.23%	1.16%	0.36%	0.0030
from	0.48%	0.51%	0.58%	0.0053
you	0.57%	0.77%	0.58%	0.0057
at	0.46%	0.42%	0.54%	0.0050
were	0.56%	0.42%	0.45%	0.0050
this	0.43%	0.54%	0.23%	0.0033
be	0.44%	0.48%	0.35%	0.0039
we	0.14%	0.23%	0.12%	0.0013
is	0.36%	0.41%	0.51%	0.0044
their	0.64%	0.25%	0.27%	0.0045

Now you might be wondering if we already had the original frequencies of each word why not just use those to calculate the mean instead of the relative frequencies?

The main reason for this is because of the difference in lengths of the texts. Valperga has 54,006 words whereas St Irvyne has 31,592, so there would naturally be a higher frequency of each word in Valperga, which makes it hard to accurately compare the values.

But by using the relative frequencies, the data is expressed as proportions, in this case a percentage, making them easier to compare and to be used for accurate calculations. Doing this also helps us to make sure that the mean reflects the authors' writing patterns rather than it being determined by the size of the texts.

Chapter 2.1

Next, we need to calculate the sample standard deviations, s, of the words' relative frequencies across the corpus.

The standard deviation, σ, is a measure used in statistics that represents how spread out a set of data values is from the mean.

The sample standard deviation is the same except it is used when we know that the data in the calculation is only a sample of the whole data set, such as how out of our corpus we are only using the values from the texts of known authors.

This can be expressed as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Where n = the total number of values, $\bar{x}$ = the mean frequency and x_i = relative frequency of an individual data value

So, to calculate the sample standard deviation for 'the', we can plug our values into the equation to get:

$$s = \sqrt{\frac{(0.0690 - 0.0673)^2 + (0.0656 - 0.0673)^2}{1}}$$

$$= 0.0024$$

If we repeat this again for the other words, we'll get something a bit like this:

	RELATIVE FREQUENCY				
WORD	VALPERGA	FRAKENSTEIN	ST IRVYNE	MEAN FREQUENCY ACROSS CORPUS	STANDARD DEVIATION
the	6.90%	5.59%	6.56%	0.0673	0.0024
of	4.39%	3.52%	4.01%	0.0420	0.0027
and	4.25%	3.97%	2.74%	0.0350	0.0106
to	2.67%	2.79%	2.82%	0.0274	0.0011
i	0.90%	3.80%	1.15%	0.0103	0.0018
a	1.97%	1.85%	1.44%	0.0170	0.0037
in	1.46%	1.51%	1.44%	0.0145	0.0002
was	1.30%	1.36%	1.16%	0.0123	0.0009
my	0.53%	2.37%	0.40%	0.0047	0.0009
that	1.25%	1.36%	0.94%	0.0110	0.0022
with	1.09%	0.89%	0.81%	0.0095	0.0020
had	0.76%	0.91%	0.79%	0.0078	0.0002
as	0.74%	0.70%	0.59%	0.0066	0.0010
by	0.77%	0.61%	0.61%	0.0069	0.0011
it	0.53%	0.73%	0.89%	0.0071	0.0026
for	0.68%	0.66%	0.68%	0.0068	0.0000
which	0.68%	0.74%	1.21%	0.0094	0.0038
on	0.59%	0.61%	0.71%	0.0065	0.0008
but	0.59%	0.92%	0.58%	0.0059	0.0001
not	0.59%	0.68%	0.64%	0.0061	0.0003
me	0.23%	1.16%	0.36%	0.0030	0.0009
from	0.48%	0.51%	0.58%	0.0053	0.0007
you	0.57%	0.77%	0.58%	0.0057	0.0001
at	0.46%	0.42%	0.54%	0.0050	0.0006
were	0.56%	0.42%	0.45%	0.0050	0.0008
this	0.43%	0.54%	0.23%	0.0033	0.0014
be	0.44%	0.48%	0.35%	0.0039	0.0006
we	0.14%	0.23%	0.12%	0.0013	0.0001
is	0.36%	0.41%	0.51%	0.0044	0.0011
their	0.64%	0.25%	0.27%	0.0045	0.0026

Chapter 2.2

Now that we've finally calculated all the means and standard deviations, we can move onto the next step which is calculating a z-score for each word in each text.

A z-score is a measure of how many standard deviations a data value is above or below the mean.

Z-scores are useful as they do something called 'standardising the data', which means that they convert the data values in a set so that they have a mean of 0 and a standard deviation of 1. If we do this with various datasets it allows us to compare data with different averages such as the most frequent words in the analysis.

But why does this work for all datasets?

A z-score can be represented by:

$$z = \frac{x - \bar{x}}{s}$$

Where $\bar{x}$ = the mean frequency, x = the relative frequency and s = the sample standard deviation

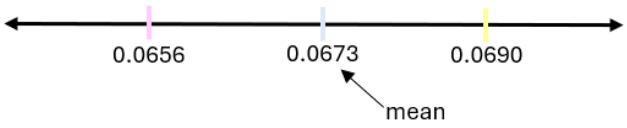
If we look at the numerator of the equation, $x - \bar{x}$, we are subtracting the mean frequency of a particular word from the relative frequency of that word in each text, which shifts all the data values by the same amount.

Therefore if $x = \bar{x}$, then

$$z = \frac{\bar{x} - \bar{x}}{s} = 0$$

So the mean shifts to zero.

For example, if you plot the relative frequencies for the word 'the' on a number line:



And then subtract the mean from each value...

You can see that the mean now shifts to zero, which you can also check by using the equation for the mean:

$$\frac{(-0.0017) + (0.0017)}{2} = 0$$

Additionally, say that , $x - \bar{x}$ were equal to the standard deviation, then when calculating the z score:

$$z = \frac{s}{s}$$

$$= 1$$

Which is why the standard deviation shifts to 1.

To calculate the z-score for 'the' in Valperga we do:

$$z = \frac{0.0690 - 0.0673}{0.0024} = 0.707$$

Repeating this with the other words, we get:

	Z-SCORES		
WORD	VALPERGA	FRAKENSTEIN	ST IRVYNE
the	0.707	-4.733	-0.707
of	0.707	-2.547	-0.707
and	0.707	0.443	-0.707
to	-0.707	0.427	0.707
i	-0.707	15.464	0.707
a	0.707	0.392	-0.707
in	0.707	2.877	-0.707
was	0.707	1.395	-0.707
my	0.707	20.500	-0.707
that	0.707	1.176	-0.707
with	0.707	-0.298	-0.707
had	-0.707	8.317	0.707
as	0.707	0.414	-0.707
by	0.707	-0.691	-0.707
it	-0.707	0.060	0.707
for	0.707	-9.682	-0.707
which	-0.707	-0.525	0.707
on	-0.707	-0.439	0.707
but	0.707	56.421	-0.707
not	-0.707	2.147	0.707
me	-0.707	9.170	0.707
from	-0.707	-0.254	0.707
you	-0.707	25.169	0.707
at	-0.707	-1.376	0.707
were	0.707	-1.014	-0.707
this	0.707	1.490	-0.707
be	0.707	1.358	-0.707
we	0.707	9.549	-0.707
is	-0.707	-0.290	0.707
their	0.707	-0.810	-0.707

Chapter 3

Now it's time for the actual distance part of the analysis!

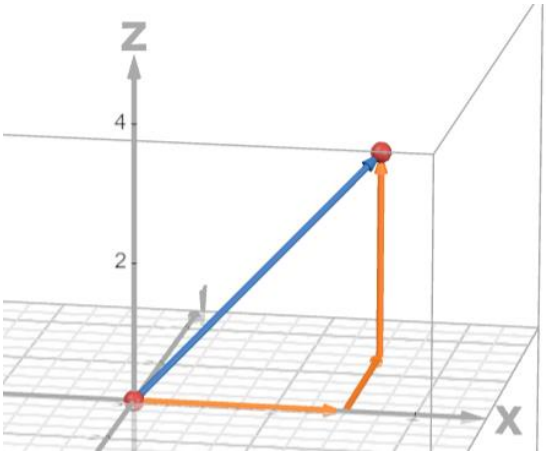
Once we have calculated the z-scores, we can say that each author has a dataset of 30 scores.

We can therefore express each author as a vector in a 30 dimensional space, where a vector is just a quantity that has both magnitude and direction. For example,

Mary Shelley would be a vector with coordinates (0.707, 0.707, 0.707, … , 0.707)

What we need to work out is the distance between the authors points and the point for Frankenstein.

To do this Burrow uses the Manhattan distance, which is total sum of the absolute differences between two points across all dimensions, where movement is limited to only the horizontal and vertical planes. Based off the journey of a taxi through the grid like streets of Manhattan!



The Manhattan distance (orange) between points in 3D space compared to the standard Euclidean distance (blue)

Say we had two vectors of n-dimensions, $A = (a_1, a_2, \dots, a_n)$, $B = (b_1, b_2, \dots, b_n)$, the Manhattan distance would be:

$$D = |a_1 - b_1| + |a_2 - b_2| + \cdots + |a_n - b_n|$$

NOTE: the | |, absolute value, is used as we want every distance to be positive, so when adding them together, they don't cancel out.

Burrow's Delta Equation can be defined as the Manhattan distance divided by the number of most frequent words:

$$\Delta_B = \frac{1}{n}\sum_{i=1}^{n}|z_i(D_1) - z_i(D_2)|$$

Where, D = the text document, n = the number of most frequent words and z_i = the z-score of an individual word in a document.

If we apply this formula to find the distance between Mary and Frankenstein, we get:

$$D = \frac{|0.707 - (-4.733)| + |0.707 - (-2{,}547)| + \cdots + |0.707 - (-0.810)|}{30}$$

$$= 6.08$$

And for Percy and Frankenstein:

$$D = \frac{|(-0.707) - (-4.733)| + |(-0.707) - (-2{,}547)| + \cdots + |(-0.707) - (-0.810)|}{30}$$

$$= 6.12$$

Epilogue

From the results of the analysis, we can see that Mary Shelley is closer to Frankenstein than Percy, suggesting that she was more likely to have written the text!

However this difference is quite small ~ 0.04, still leaving a bit of uncertainty. This is possibly because Burrows Method works best when 100-150 of the most frequent words are used instead. To further increase the accuracy of the analysis we could use a much larger corpus with more than one text from each author to develop a better authorial fingerprint.

We could also use different stylometric methods to further back the analysis such as the imposter algorithm, or even develop neural networks using computer softwares!

Overall, stylometry truly is an interestingly diverse field, that has helped to breathe life into the world of literature, fusing the two opposing fields of maths and English into something great!