3-21-2011

Computational Stylometry: An Interdisciplinary Project

Abby Miller
Johnson & Wales University - Providence, alm860@jwu.edu

Taylor Horn-Speck
Johnson & Wales University - Providence, trh276@jwu.edu

Blair Mondino
Johnson & Wales University - Providence, bmm623@jwu.edu

John "J.C." White
Johnson & Wales University - Providence, jcw723@jwu.edu

Computational Stylometry

Campus Reads: No Impact Man, Colin Beavan

An Interdisciplinary Project: Professor Eileen Medeiros' Freshman (Honors) English Composition Class and Professor Richard Cooney's Statistics Class

Contributions by: Abby Miller, Taylor Horn-Speck, Blair Mondino, and John "J.C." White

Computational Stylometry

Introduction

Until the 1960's, it was unknown who wrote many of the Federalist Papers of the late 1700's. The 85 Federalist Papers were written to persuade the citizens of the State of New York to ratify the Constitution. They were written under the pseudonym "Publius" which involved three political figures: Alexander Hamilton, James Madison and John Jay. Twelve of these papers were called "disputed" because historians were unsure whether they were written by Alexander Hamilton or James Madison. In the 1960's, two statisticians, Frederick Mosteller of Harvard University and David Wallace of The University of Chicago applied a research method called Computational Stylometry and provided statistical evidence that strongly suggests that the disputed papers were Madisons.

Method

Stylometry is the study of a linguistic style as it normally applies to written language. More simply, it is the study of certain aspects of style. While this is normally applied to language, stylometry has been used in other applications like art and music. However, for the study conducted by both Professor Cooney's class and Professor Medeiros' classes, we focused on Computational Stylometry. Using this method, we applied math and counting with literary elements to help determine the authorship of a piece of writing.

Computational Stylometry is used to help identify authors by their subconscious elements of style. It is studying a style that is normally unnoticeable because the words and techniques are commonly used. This is where the numbers and math come into play. Perhaps by counting these nuances in style, we can identify an author's unintentional style and attribute anonymous pieces to their rightful owners. According to Holmes, McEnery and Oates, these nuances in writing are called discriminators. Computational Stylometry is not an infallible way to identify an author because styles and techniques may change over a period of time. However, they have been shown to aid in identifying authors of anonymous pieces, identifying

plagiarism, and resolve issues of disputed ownership as seen with the Federalist Papers.

Over the course of English 1930 during the fall term, we studied many different types of methods of research and sources. When given the chance to collaborate with another class studying a new style of research, the students accepted the challenge. The English 1930 class was given four documents to review. In each document, students had to count the number of occurrences of the words "and", "but" and "by" in the four documents. Professor Cooney's Math 2001 class counted the number of occurrences of the words "for", "to" and "from".

Results

In my class, we collected the data from each piece and created a chart. This is when we handed off our data to Professor Cooney and his students to organize, analyze, display and make a conclusion as to which of the three manuscripts (whose authorship was unknown) is actually that of Colin Beavan, No Impact Man, "Chapter 3".

In Professor Cooney's Statistics class the type of reasoning applied to the data collected in this research project is inferential, not deductive. Our conclusion does not prove findings with certainty, instead we will report what is very likely.

The six words chosen were independent of the context of the manuscript selections being studied. The rate occurrence of each word was described as a relative frequency at a rate per thousand. A relative frequency distribution table was created to compare results. A relative frequency histogram was constructed to graphically display the results. Residuals were calculated to determine deviations from expected values, and from these residuals a conclusion was made.

Relative Frequency Distribution (Words per thousand)

Colin Beavan, "No Impact Man"

"Chapter 3" (Known)

Key word	*Total words*	*# occurrences*	*Relative frequency*
"and"	5,136	109	109/5136 = .021
"but"	5,136	20	20/5136 = .004
"by"	5,136	6	6/5136 = .001
"for"	5,136	46	46/5136 = .009
"to"	5,136	154	154/5136 = .030
"from"	5,136	10	10/5136 = .002

"Bird Leg" (Unknown author)

Key word	*Total words*	*# occurrences*	*Relative frequency*
"and"	2,646	70	70/2646 = .026
"but"	2,646	9	9/2646 = .003
"by"	2,646	5	5/2646 = .002
"for"	2,646	15	15/2646 = .006
"to"	2,646	58	58/2646 = .022
"from"	2,646	6	6/2646 = .002

"Chapter 4" (Unknown author)

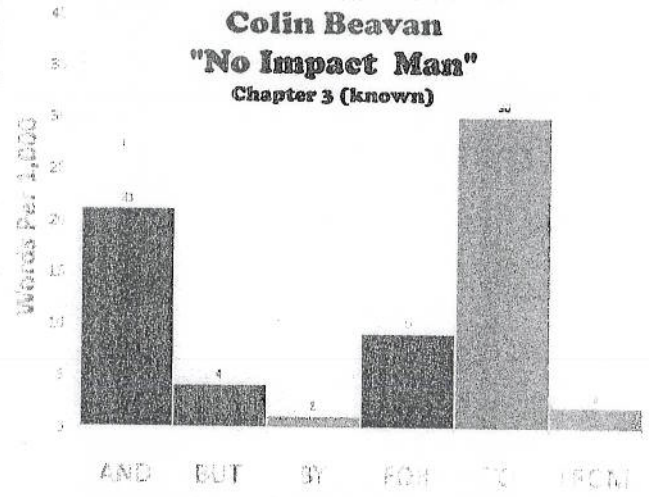
Key word	*Total words*	*# occurrences*	*Relative frequency*
"and"	2,749	75	75/2749 = .027
"but"	2,749	4	4/2749 = .001
"by"	2,749	6	6/2749 = .002
"for"	2,749	22	22/2749 = .008
"to"	2,749	72	72/2749 = .026
"from"	2,749	5	5/2749 = .002

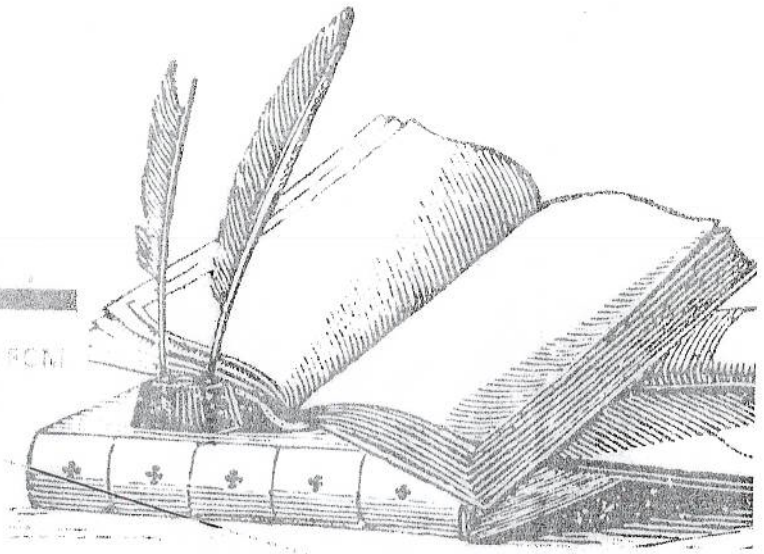
"My Parents" (Unknown author")

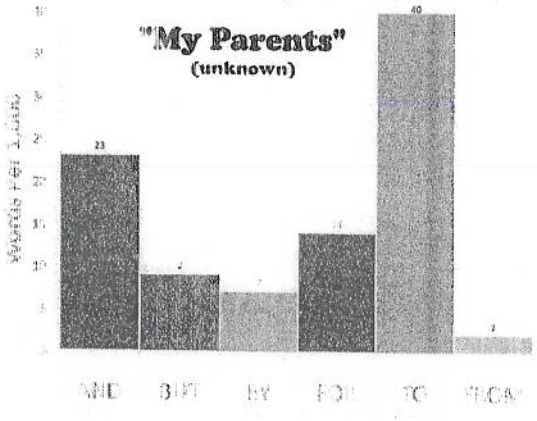
Key word	*Total words*	*# occurrences*	*Relative frequency*
"and"	1,368	31	31/1368 = .023
"but"	1,368	12	12/1368 = .009
"by"	1,368	10	10/1368 = .007
"for"	1,368	19	19/1368 = .014
"to"	1,368	55	55/1368 = .040
"from"	1,368	3	3/1368 = .002

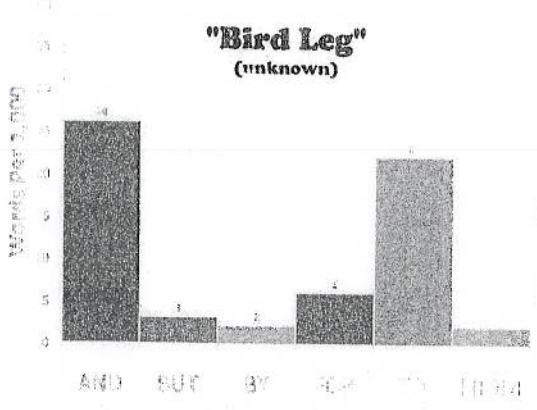
When viewing the display presentation, you will observe that all three relative frequency histogram graphs of the unknown authors, which represent the relative frequencies of the six common words, appear to be very similar to the relative frequency graph of Colin Beavan's No Impact Man, "Chapter 3". A method used to determine any significant differences would be to take a look at the *residuals*. A residual is the difference between the observed value and the expected value. For each key word, the relative frequency (per thousand words) of Colin Beavan's No Impact Man is known. If the relative frequency (per thousand words) of the same word from each of the three unknown authors is subtracted from Colin Beavan's known relative frequency, then the value of the residual can be determined. Some of the differences will be more than the expected value (+), and some of the differences will be less than the expected value (-). If each of these differences is squared, then the products all become positive (+). To find which unknown author model will be the "Best Fit" to the actual model of Colin Beavan's No Impact Man, we can calculate the the sum of the squares of the differences (residuals). The model that has the smallest sum will be the best fit and match for the authorship of this piece of literature.

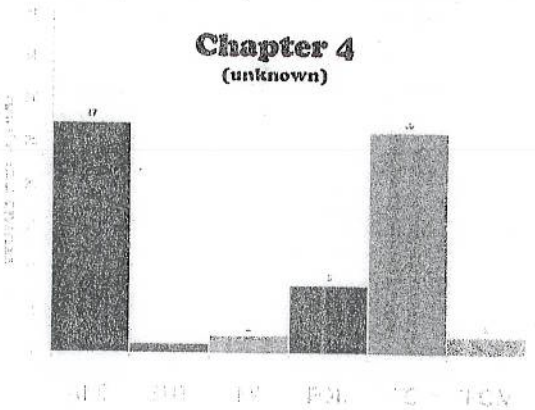
"Computational Stylometry"











By John White

"Bird leg" (Unknown author)

Relative frequency (Words per thousand)

Key word	***Residuals (Observed - Expected)***	***$(Residual)^2$***
"and"	.026 - .021 = .005	.000025
"but"	.003 - .004 = -.001	.000001
"by"	.002 - .001 = .001	.000001
"for'	.006 - .009 = -.003	.000009
"to"	.022 - .030 = -.008	.000064
"from"	.002 - .002 = .000	+ .000000
		.000100

"Chapter 4" (Unknown author)

Relative frequency (Words per thousand)

Key word	***Residuals (Observed - Expected)***	***$(Residual)^2$***
"and"	.027 - .021 = .006	.000036
"but"	.001 - .004 = -.003	.000009
"by"	.002 - .001 = .001	.000001
"for"	.008 - .009 = -.001	.000001
"to"	.026 - .030 = -.004	.000016
"from"	.002 - .002 = .000	+ .000000
		.000063

"My parents" (Unknown author)

Relative frequency (Words per thousand)

Key word	***Residuals (Observed - Expected)***	***$(Residual)^2$***
"and"	.023 - .021 = .002	.000004
"but"	.009 - .004 = .005	.000025
"by"	.007 - .001 = .006	.000036
"for"	.014 - .009 = .005	.000025
"to"	.040 - .030 = .010	.000100
"From"	.002 - .002 = .000	+ .000000
		.000190

The calculations reveal that the sum of the squares of the residuals for each manuscript of the respective unknown authors are:

***"Bird leg"* is .000100**

***"Chapter 4"* is .000063**

***"My parents"* is .000190**

The model that has the *smallest* sum of the squares of the residuals is *"Chapter 4"*. Using *Computational Stylometry*, we can statistically predict that the author of the manuscript labeled *Chapter 4"* is Colin Beavan.

Colin Beavan *is*, in fact, the author of this piece of literature!

Works Cited

Beavan, Colin. ***No Impact Man: The Adventures of a Guilty Liberal Who Attempts to Save the Planet and the Discoveries He Makes About Himself and Our Way of Life in the Process.*** **New York: Picador, 2009.**

Operation Jedburgh: D-Day and America's First Shadow War. **New York: Viking, 2006.**

Holmes, D., T. McEnery, and M. Oates. "Stylometry and Authorship." ***Docstoc - Documents, Templates, Forms, Ebooks, Papers & Presentations.*** **2010. Web. 20 Mar. 2011.**

Kotlowitz, Alex. ***There Are No Children Here: The Story of Two Boys Growing Up In the Other America.*** **New York: Doubleday, 1991.**

Thomas, Lewis. ***Late Night Thoughts on Listening to Mahler's Ninth Symphony.*** **New York: Viking, 1980.**

Usiskin, Zalmin. ***Functions, Statistics, and Trigonometry.*** **Glenville, Illinois: Scott, Foresman, 1992.**