An Introduction to Stylometric Analysis

Nathaniel Latta

Abstract:

This paper gives an overview of the field of Stylometry. It starts with an introduction explaining how Stylometry works and what makes it important. Next, I will investigate several papers to see the kinds of tools that are being researched as ways to expand Stylometric analysis. After looking at a few papers I will give an overview of three platforms for performing Stylometric analysis: Signature, JGAAP, and Stylo. The paper concludes with an example of a stylometric analysis using these platforms.

Introduction:

Stylometry is the study of linguistic style, this means looking at patterns in language to outline rules or characteristics of the subject of study. Stylometry is often used to determine authorship to anonymous manuscripts, but the principles of Stylometry can also be used in other areas such as the analysis of music. Stylometrists search for a clear measurable attribute that can be used to draw definitive conclusions about an author. It is possible, however, that such a feature does not exist. However, as the search progresses, researchers find statistical attributes and develop tools that, when used in conjunction with each other, allow very strong conclusions to be drawn in authorship disputes.

The basics of stylometry were established by the Polish philosopher Wincenty Lutoslawski in 1890. In the past, this discipline was very limited because humans had to manually perform all of the data analysis on the texts they were studying. With the advent of computers, however, researchers are freed from data analysis roles to focus on the theory underlying the data relationships. Stylometry gained popularity from renaissance drama authorship questions. This interest in determining authorship in renaissance works helped to establish stylometry's credibility in the area of authorship attributions. A simple concept to begin with in stylometry is the idea of the writer invariant. This is a writing characteristic that belongs to a specific author. The idea is that, the author of a document has features of his or her writing which are consistent across all documents written and that all authors have their own unique invariants. Hence, identification of this attribute can help to determine who wrote a document. It is agreed that these types of traits do exist between authors, but trouble exists in identifying which aspects of the text should or should not contribute to this invariant.

There are three different categories which general stylometry can be broken into. These are: authorship attribution, characterization, and verification. Authorship attribution is determining the correct author out of a small group of authors. Authorship Characterization involves determining the physical characteristics of an author such as age, gender, or race. Authorship verification is determining if a document was written by a specific individual. The differences between verification and attribution are very nuanced. If the reader wishes to learn more about the differences, the paper *Authorship verification for short messages using*

Stylometry (Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on. IEEE, 2013.) by Brocardo is a good start. Most research in stylometry falls into the category of attribution, with verification being the least developed, having been studied almost exclusively in the context of plagiarism. Stylometric tests are generally targeted at long documents being unreliable on short documents. A long document is a fairly general term often dependent upon the study. Short documents are more easily described as documents one or two paragraphs in length. The ability of Stylometry to analyse these short documents is a developing interest area. Stylometry's success on works such as novels and plays has raised interest in its merits in modern contexts where online document length is much shorter. Examples of the focus of modern stylometry are email or forum posts. These document sources provide additional challenges beyond their length because they are often poorly structured or written. This is an important area of development because reliable stylometric tools for short online documents can be useful in criminal cases giving law enforcement more tools to prosecute criminals. These new applications for stylometry have raised privacy concerns that continue to be discussed. If the reader is interested in learning more about the privacy concerns around Stylometry, Mike Brennan's works (https://www.youtube.com/watch?v=STKwpNYzWis) or (Brennan, Michael, Sadia Afroz, and Rachel Greenstadt. "Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity." *ACM Transactions on Information and System Security (TISSEC)* 15.3 (2012): 12.) are good places to start.

Burrow's Delta Metric:

The first two papers that are helpful in gaining an understanding of the existing field are a pair of papers by John Burrows: *Questions of Authorship: Attribution and Beyond*, and *"Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship*. Both papers revolve around the Delta metric he has developed for designating likely authorship but the first paper focuses more on the meaning and impact of this metric while the second paper discusses the detailed evidence supporting Delta as a metric. Delta, put simply, is a measurement of the mean of the positive and negative differences rendered as absolute divergences. This means that Delta is a normalized measurement of difference.

In *Questions of Authorship*, John Burrows explains that the advent of computers allows us to gather more subtle differences in texts as well as measure them more swiftly and accurately than was possible with manual analysis. Current tests behave by asking questions like: is this text closer to author *X* or *Y*? That is, a text is grouped with an author and measured in terms of similarity to that author. We must always do additional tests because we know nothing about the possibilities of a third author. It is a common mistake to think that because author *X* is least unlike author *Y*, they then share meaningful similarities. In such tests it is more important to consider patterns that the texts fall into rather than a text's similarities to an author. To reiterate,the grouping of texts around author *X* is a grouping of least different. Therefore, it is dangerous to draw conclusions about traits the texts share. Because of this, patterns present in a text are more informative than trying to find similarities. When looking for a holy grail metric for stylometry, the author thinks we will not find one within statistical analysis because, by nature, statistics deals with probabilities

not certainties. However, strong probabilities when combined with judgement are incredibly useful.

Burrow's second paper discussing the particulars of the Delta test begins by explaining how the methods stylometrists have developed are best fitted for 'closed games,' where we are working with a small set of authors and a large piece of text. If, however, we find ourselves in situations where we do not have outside evidence to use to determine the author, our tools are much less effective. It has been suggested that we should pass over these situations and only analyze a small number of authors' texts at a time, but this Delta procedure helps to turn an 'open game' into a 'closed game'. Currently stylistic signature is measured by many tiny strokes. Many small, weak indicators of authorship are more reliable than a few strong ones. These methods have many subtle intricacies that are useful in direct comparisons but are unhelpful in creating a ranking. To address this area with a distinct lack of tools, the Delta score seeks to be an expression of pure difference. The first step is to make a frequency-hierarchy for the most common words in a large group of texts from authors relevant to the investigation. (Spelling and contraction should be standardized. When working with word frequencies it is not an important feature). The next step is to take the positive and negative measures of distance and transform them into absolute differences. These positive and negative measures are things such as how each document's word values, or *z*-score, compares to the group as a whole. Delta score is defined as: "the mean of the absolute differences between the *z*-scores for a set of word-variables in a given text-group and the *z*-scores for the same set of word-variables in a target text" The distinction of *z*-scores and not total percents is very important. The paper discusses results of their study which were quite positive, being able to reliably identify the correct author from an unknown text. It concludes that this method can be used on a large group of works and authors to rank texts by their most likely author to a degree significantly more reliable and precise than what would be expected of random chance.

Greek Stylometry:

Next, we shall move on to look at some cases of Stylometry applied to ancient Greek. The paper *The New Stylometry: a one-word test of authorship for Greek writers* by Michaelson and Morton looks at a technique used to help determine authorship of ancient Greek texts. To do this, the authors of the paper propose that certain characteristics of written works exist that are common between all Greek authors. If one of these characteristics can be identified, it could be used to identify the author if the frequency of these occurrences can be shown to vary from author to author as well as remain consistent across a single author's works. Such an event must be very precisely defined. The event must occur frequently in order to do a proper analysis: five times minimum for a section, but, in practice, many more than five are needed. A section refers to the size of the text being analyzed for example to analyse a page of text this method requires at least five occurrences on the single page. In the paper, they walk through an example using the occurrence of a particular Greek word which they determine meets the strict requirements of the method. The article notes that for the purpose of this analysis the difference between occurrence and usage of the word is not important. Statistical analysis can be used to measure if the difference between expected frequency of a word and the actual frequency of

the word can be attributed to chance. The paper walks through an example that shows that different authors do, in fact, use the chosen words at reliably different frequency. A single author's usage of the word remains consistent over multiple works. This means that, when compared to a different author who uses the word at a different frequency, the work can be attributed to the first author. The paper concludes by acknowledging that this tool cannot be used for sensitive checks of authorship, but can be very useful in combination with other methods. A particular example where the method would break down is if two different authors use the chosen word or feature at the same frequency. In such a situation, the authorship of the work could be narrowed down to authors that employ that frequency but it could not identify a singular author.

The next paper is also a stylometric analysis identified for Greek authors, W. C. Wake's *Sentence length distributions of Greek authors* explains that Greek relies upon three punctuation marks: the full stop (.); the Greek colon placed above the line (a period floating in the middle of the line); and the interrogation mark (;). Greek punctuation is a modern invention, and there is no guarantee that the punctuation added matches the original author's intention. However, statistical period (sentence) length can be determined from sentences with a clear ending. This helps us in general cases as language generally follows specific patterns. However, in objectionable circumstances, statistical reasonings can be suspect. It is likely that the original manuscript of an author would have had natural breaks to reflect trains of thought. Ancient scribes intimately familiar with the spoken language would have found it intuitive to punctuate based upon sentence structure. Most copies of ancient writings that exist already have had punctuation added in this manner. It is likely that these punctuations would reflect those of the original due to their proximity to it. This means that we can safely use the punctuations provided as we have no good reason to suspect that the length of these sentences has been altered. When selecting sentences to analyze, the authors do not analyze sentences where there is significant variation of length between copies. This results in less than 2% of sentences being discarded. Wake concludes that sentence length distributions display a consistency between an author's works, and that this method offers a soft indicator of authorship.

These two papers regarding Greek authorship have shown us the point illustrated by Burrows: that stylometric analysis consists of many soft indicators of authorship that all work together to predict the most likely author. With these ideas in mind, we can move on to look at some tools for conducting stylometric analysis using some of the tools that have been discovered.

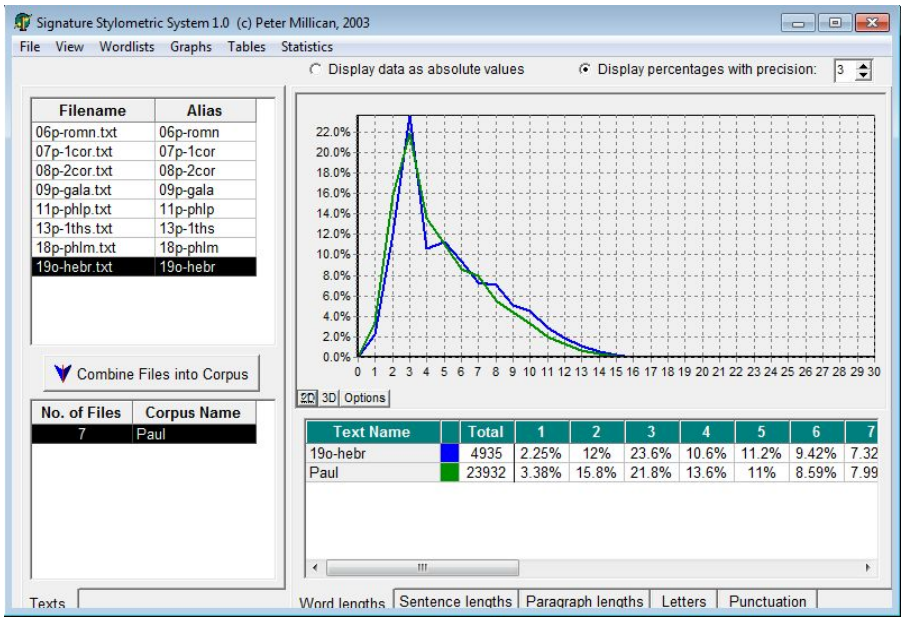
Software:

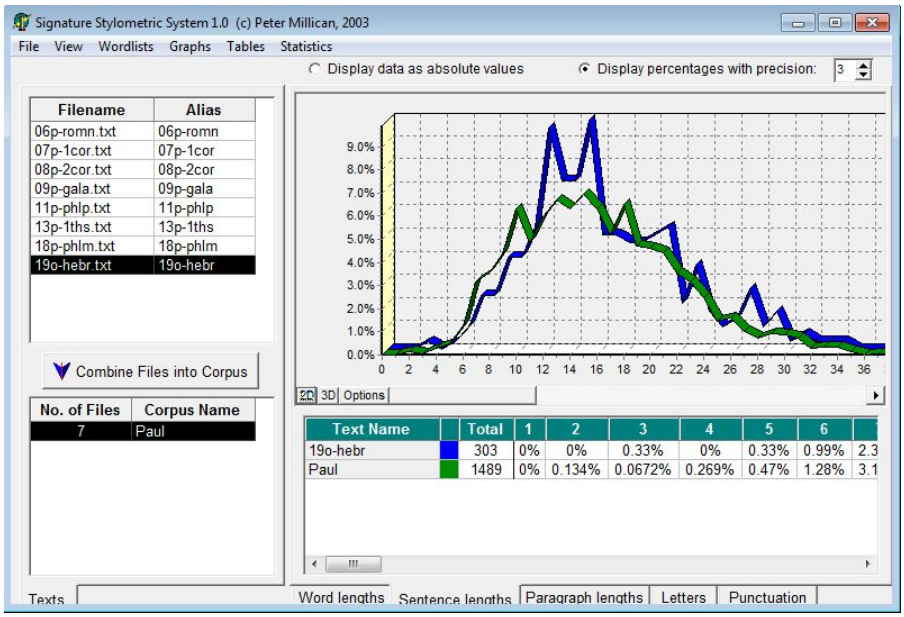
Here, I will discuss three prominent tool packages for do-it-yourself stylometry: *Signature Stylometric System, Java Graphical Authorship Attribution Program* (JGAAP) and the R library *Stylo*. I will also discuss the advantages and disadvantages of each.

Signature Stylometric system:

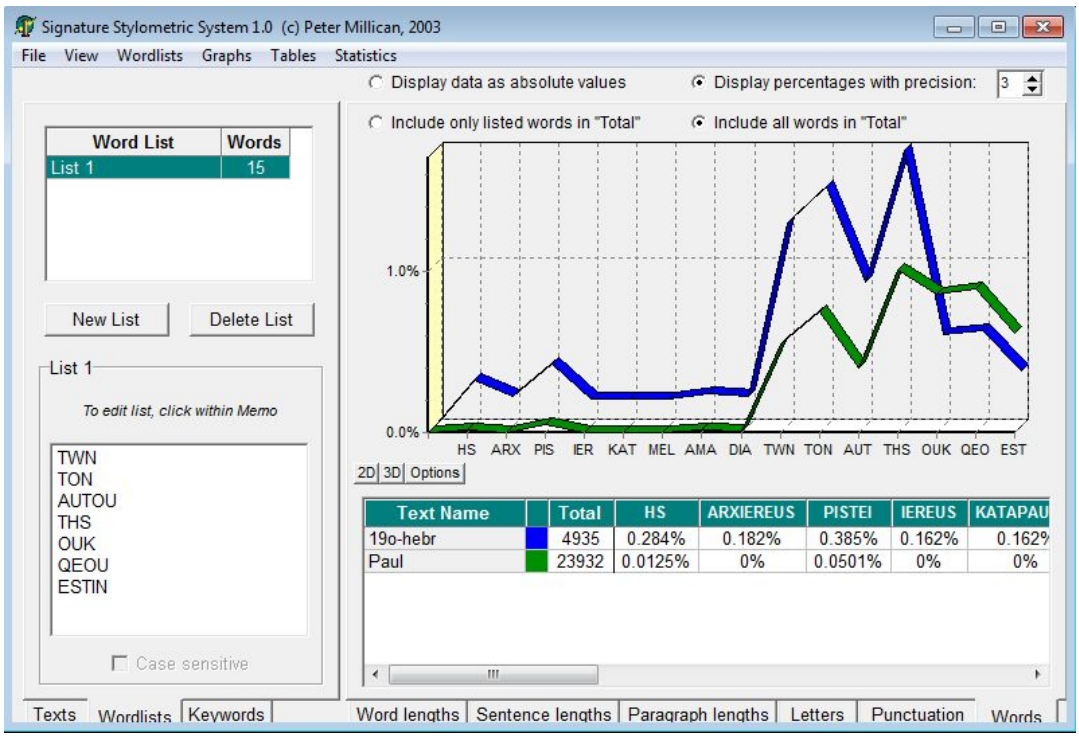
Signature is the most basic, and offers simple ways to analyze documents by directly comparing properties against each other. Signature allows you to compare word length, sentence length, paragraph length and punctuation and letter frequency. Based on these attributes, Signature generates a graph comparing numerous documents with the option of measuring values as percents for standardization or doing a chi-square test on two documents. Signature also has the ability to generate keywords and graph the occurrences of selected words. This is less useful than it sounds though—there is no documentation included explaining how keywords are ordered and signature can only graph user-selected words leaving the user with no guidelines on how to create meaningful word comparisons.

These examples compare word, and then sentence lengths of some Biblical works of the Apostle Paul (Romans, 1 and 2 Corinthians, Galatians, Philippians, 1 Thessalonians, Philemon) against the book of Hebrews whose authorship is contested.

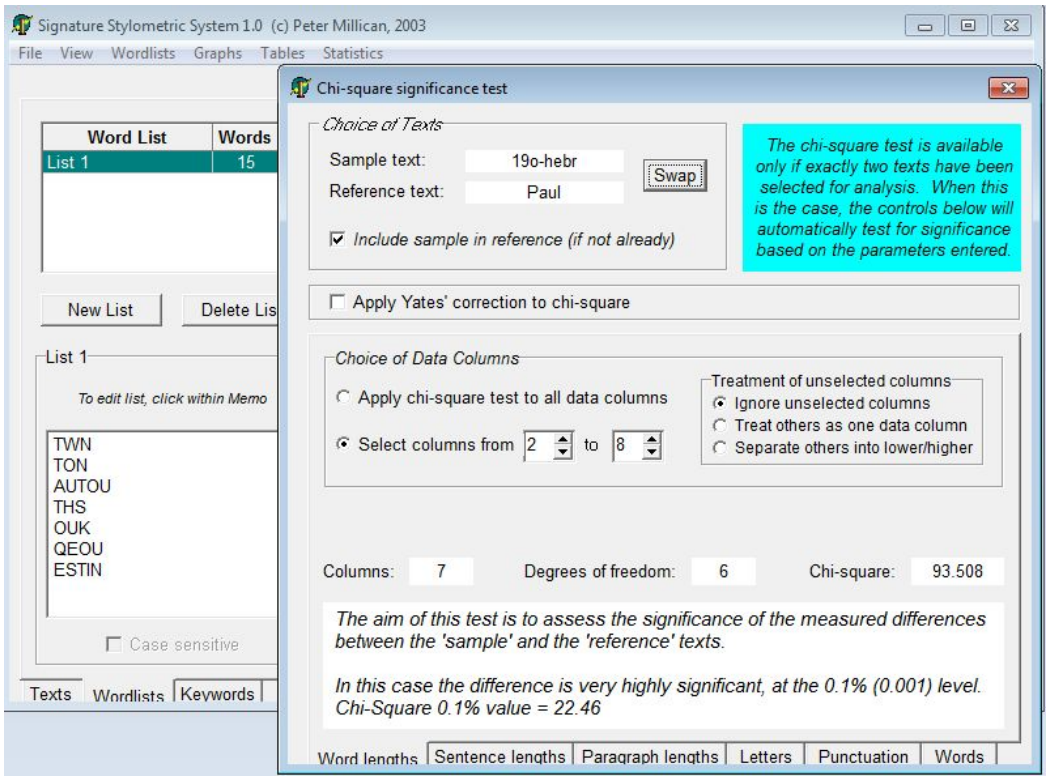




Signature can generate a list of keywords. Selected words from this list can be compared against each other as follows:



For each category, Signature can take two works and directly compare them against each other by conducting statistical tests. Here is the Chi-squared test for word length:



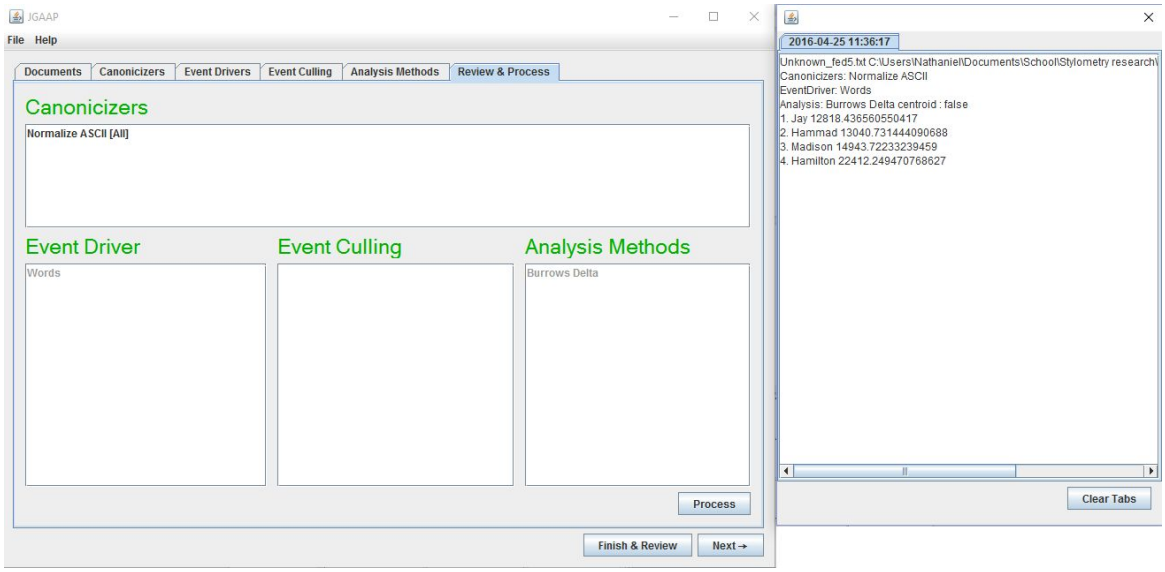
From these images, we can see that the usage of the selected words in Hebrews versus the other works of Paul are significantly different. Because, however, Signature offers no guidance to select the words to compare, this is inconclusive; there is no argument for the validity of the use of those words in a comparison. A proper way to create a word comparison might be to generate a most frequent word list for the two groups and compare those words against each other. However, Signature offers no way to do this itself. Such a comparison could be made with the help of the third program used: Stylo, which presents the user with a list of the most frequent words.

JGAAP:

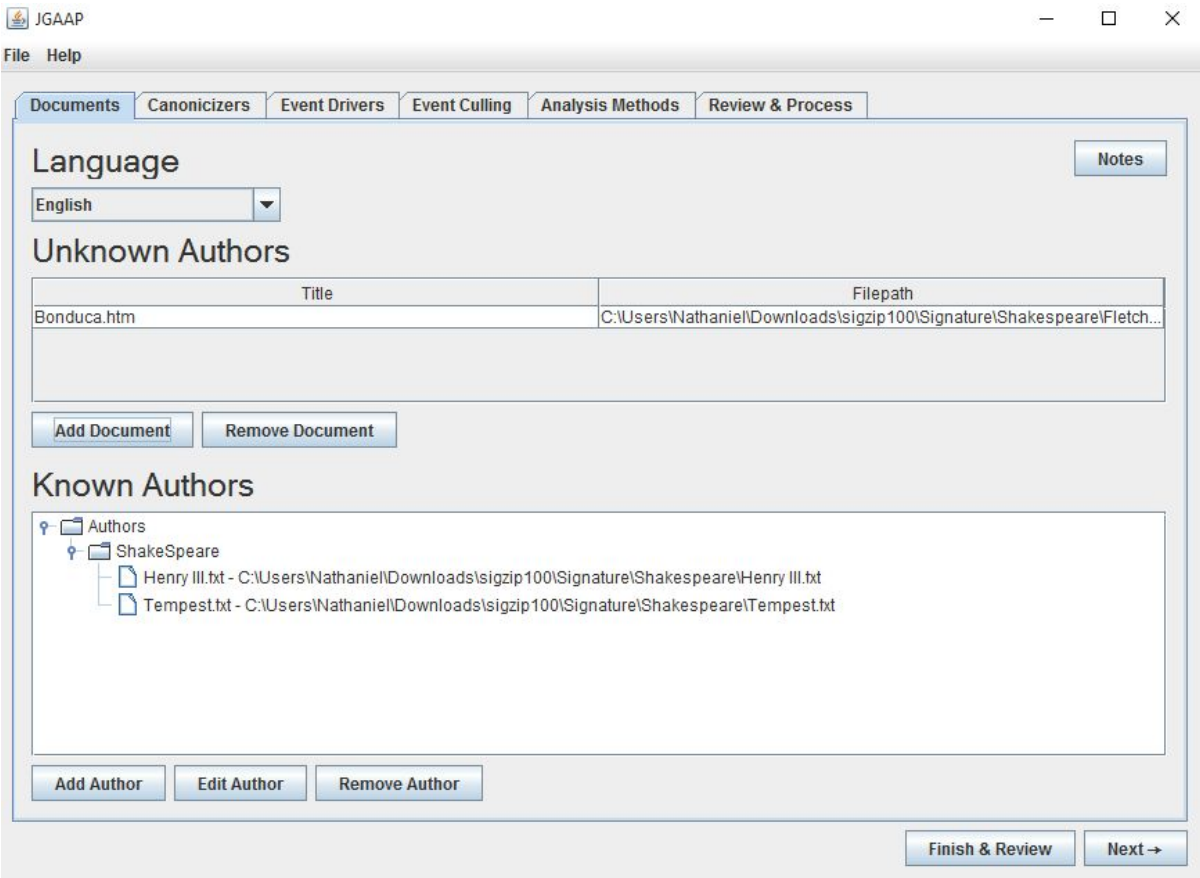
On the other hand, JGAAP allows the user much more precise control over the type of analysis being done. This could mean doing an analysis based off of the rare words in the documents, word frequency, syllables per word etc. JGAAP has a wide range of what it calls *events* that are used to compare texts. It also offers many statistical tools to customize what type of analysis is being done, such as the Burrows Delta test discussed above. All of these things lead to a powerful tool with a lot of potential, but in an untrained hand, the precise nature of stylometric analysis makes it very difficult to conduct meaningful analysis. JGAAP also does not offer a way to view what the words that drive its events are. For example, if you were to use JGAAP to do an analysis on the *n* most infrequent words of a text, JGAAP would only report the statistics generated from using these words. The words themselves are never revealed to the user.

The following image shows the way that JGAAP processes and displays its information. There is a wide range of possible entries for the four categories. “Canonicizers” are ways to standardize the text of the documents being analysed. “Event Drivers” are ways to change the documents to expose characteristics (Event Drivers could be used to focus

the analysis on the sentence lengths of a document or maybe the usage of words beginning with vowels). “Event Culling” is used to select which parts of the data to analyze (Analyze *N* events with highest standard deviation). Finally “Analysis Methods” allows the selection of the statistical tools used to analyse the data (such as the aforementioned Delta metric). The left part of the image shows the conditions set for the analysis and the right shows the result window once the analysis is complete.

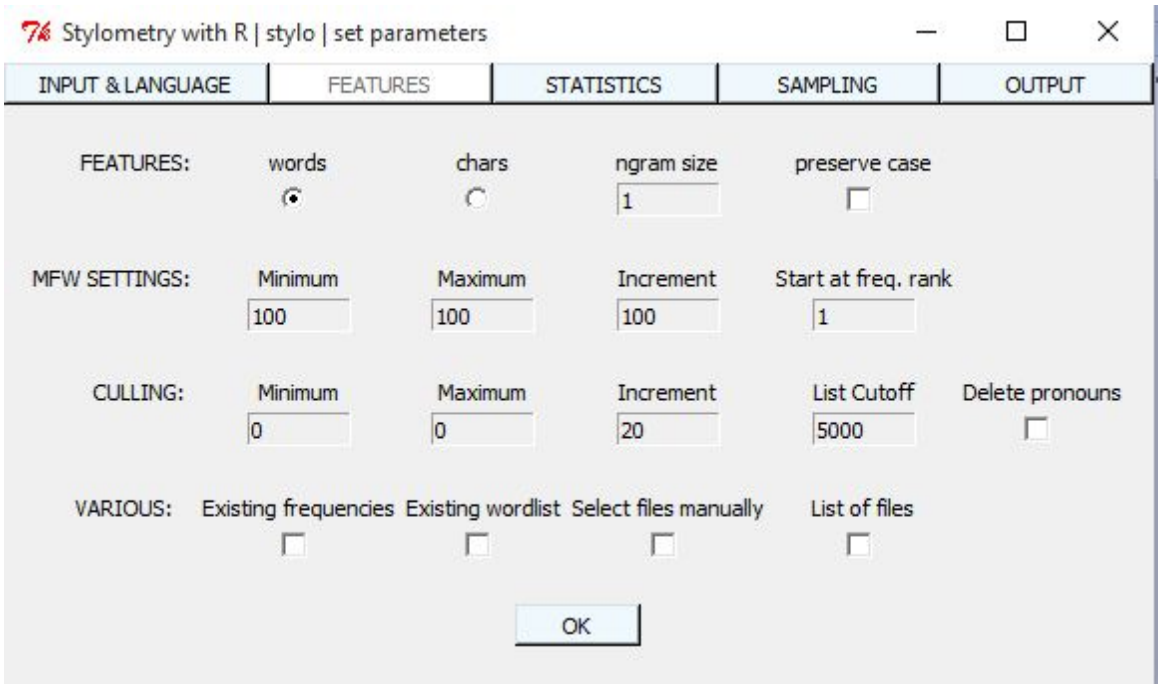


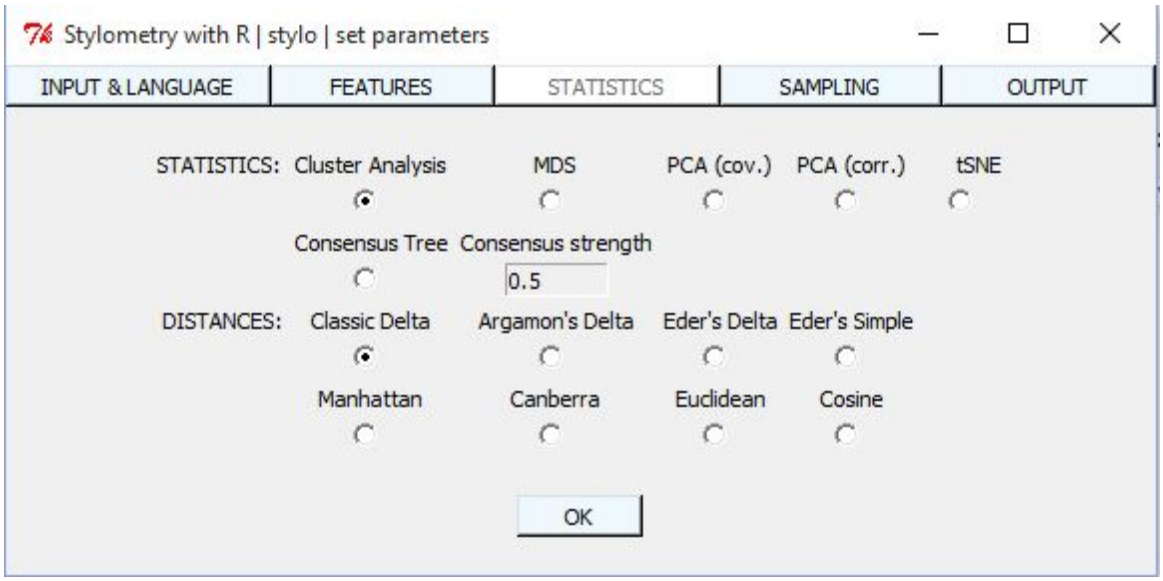
JGAAP’s input is a collection of texts belonging to specific authors and an unknown text field. This example compares two of Shakespeare’s plays against a third unknown.



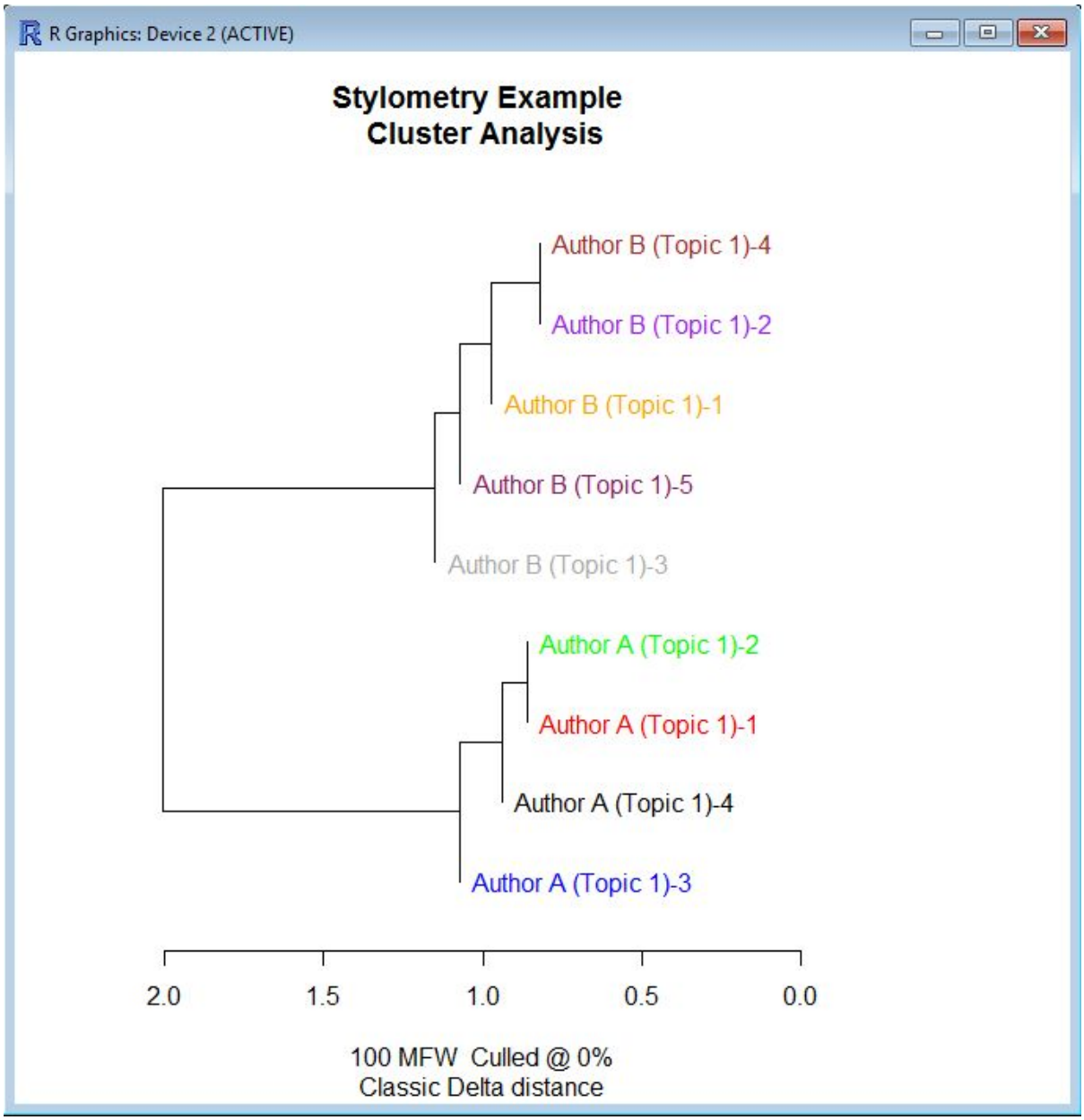
Stylo:

The third program to facilitate stylometry that I investigated was the R stylometry package: Stylo. This toolset operates within the R programing environment and features a GUI for users who do not wish to use the programming support of R. For users that wish to take advantage of R's power, Stylo includes additional functions and ways to incorporate them all into other programs. Stylo also features support for machine learning which emphasises the scope of the toolset. Because Stylo works within R, setting it up and using the basic features becomes slightly more difficult for users unversed in programing languages. Once setup is complete, however, Stylo includes a GUI to aid amature analysis. The stylo manual was very comprehensive and walked through all of the different components and how they work. The following two images show two of the tabs of the stylo() function GUI:

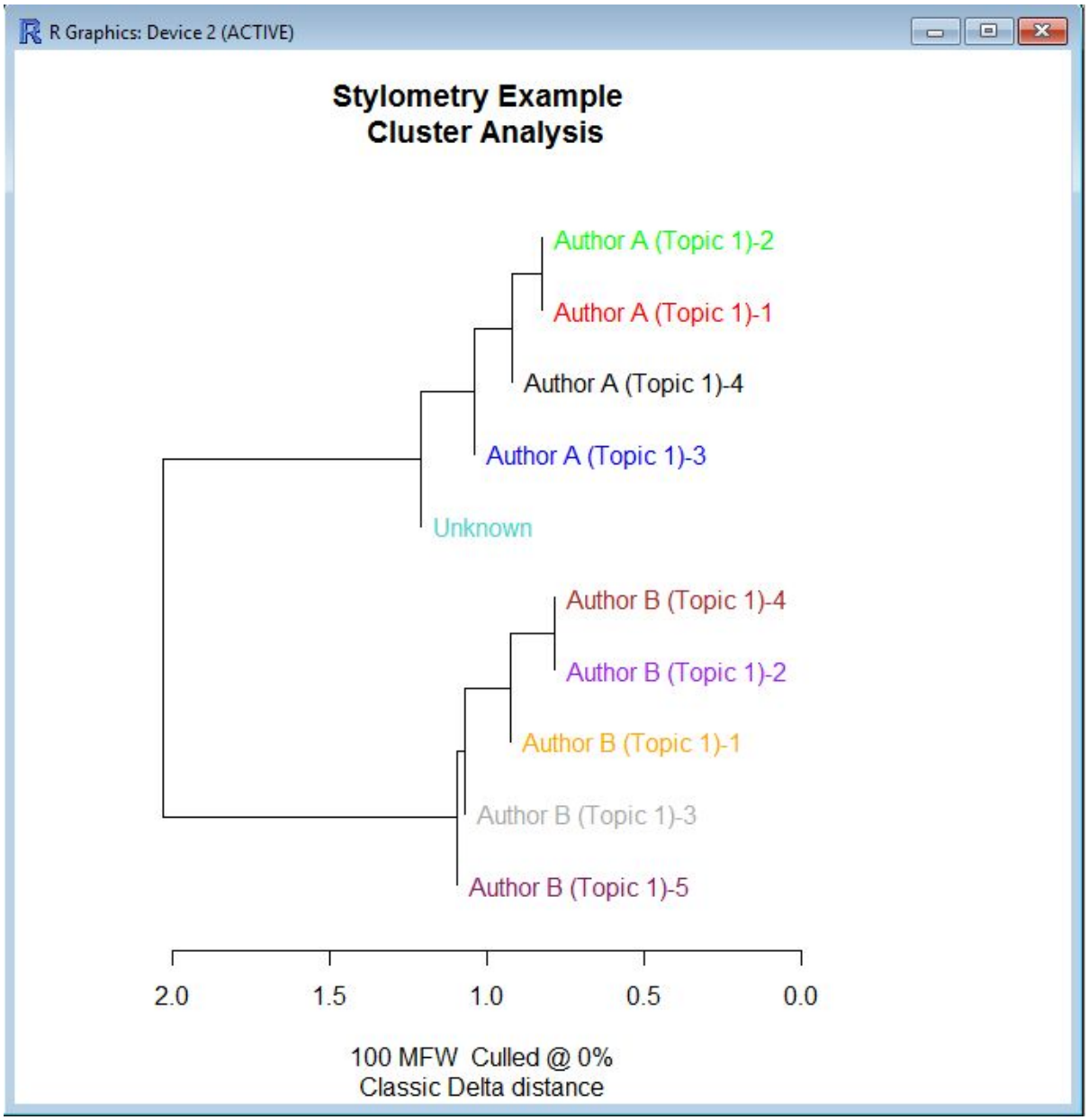




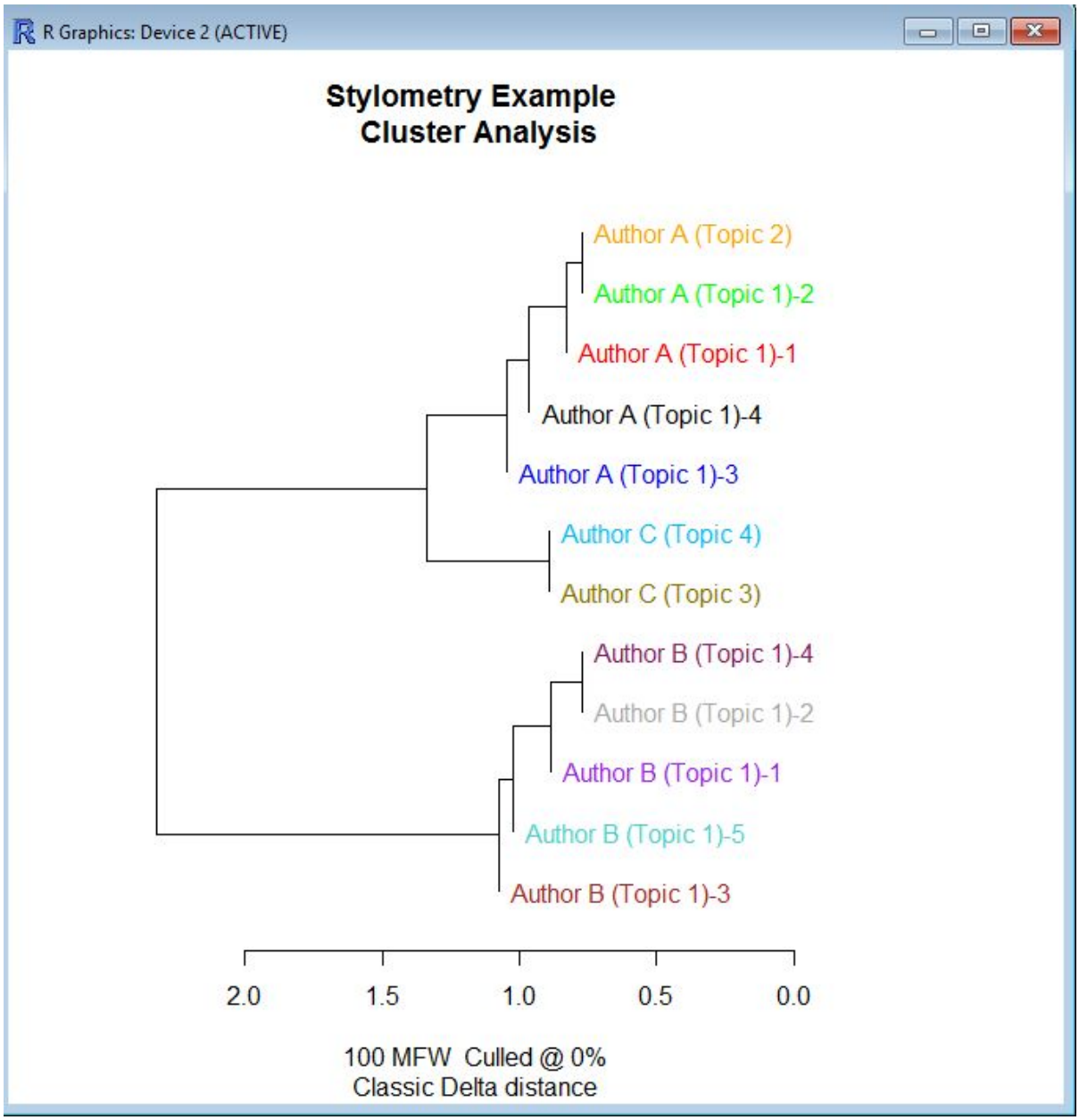
While all of the options are explained in the manual, new users can jump right into a basic analysis by using the defaults. The next image shows the results of running an analysis using the default settings. In addition to the cluster analysis show in the image, Stylo also documents the most frequent word lists that it uses for analysis and creates an indexed table of them for easy access. These most frequent word lists are the foundation upon which stylo will perform this basic analysis. For the first time user the addition of these word lists is most useful in satisfying curiosity as any analysis based off of the word lists should have deliberate motivation.



We can see that Stylo was able to group these documents according to author. This cluster analysis groups documents by similarity. Because the documents are labeled by author, we can see that this cluster analysis also has grouped the documents in such a way that we can infer authorship had one of the documents been an unknown. However, this outlines an important trap to avoid. This analysis does not tell us that each main branch is a different author. For example, consider this cluster analysis with an unknown document.



It would be incorrect to conclude that the unknown document belongs to Author *A* since this analysis only tells us that the unknown author is more similar to A than B. This unknown author is indeed a third author, "*C,*" and it is worth noting that looking at only the Author *A* branch that the unknown document is in a branch by itself separate from all of Author *A*'s works. Another document from Author *C* confirms this behaviour.



This is why it is important not to read definitive conclusions into the analysis, and to conduct multiple tests. A next step from this cluster analysis might be to compare a document from author *C* (unknown) to the most similar from Author *A* and get a numeric measurement of the similarity (JGAAP would be useful for this) and a measurement for documents between Author *A* and compare the magnitude of the difference. If the difference between Author *C* (topic 4) and Author *A* (Topic 1)-3 is significantly larger than the difference between Author *A* (topic 1)-3 and Author A (topic 1)-4 then the argument for Author *C* being a unique author separate from *A* is strengthened. This example shows us the way that many small indicators allow us to make stronger conclusions about the identity of an author.

Conclusion:

We can see from this overview that Stylometry is a field where researchers look for an authorship attribute and instead find smaller indicators that alone cannot characterize an author. At an entry level, stylometry is performing analysis on data and applying a human touch to the interpretation of that data. As the analysis gets more and more thorough and complex, professionals move away from this judgement process and end up with tools that allow them to make confident claims about Authorship. These tools take the form of statistical tests that tell us about a document (such as Delta) as well as software to facilitate the combination of multiple statistical tests (JGAAP, and Stylo). Among the three different software tools investigated only Stylo had clear functionality and support for a user to perform analysis without extensive background reading. For this paper the next step in creating understandable beginner stylometric analysis would be to create examples using multiple statistical tests and multiple different documents and authors to develop a reliable set of analysis tools to make a confident judgement about authorship.

Works Cited:

"Stylometry." *Wikipedia*. Wikimedia Foundation, n.d. Web. 11 Mar. 2016.

Burrows, John. "'Delta': A Measure Of Stylistic Difference And A Guide To Likely Authorship." *Literary And Linguistic Computing: Journal Of The Association For Literary And Linguistic Computing* 17.3 (2002): 267-287. *MLA International Bibliography*. Web. 11 Mar. 2016

Burrows, John. "Questions of Authorship: Attribution and Beyond: A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York." *Computers and the Humanities* 2003: 5. *JSTOR Journals*. Web. 11 Mar. 2016.

Michaelson, S., and A. Q. Morton. "The New Stylometry: A One-Word Test of Authorship for Greek Writers." *The Classical Quarterly* 1972: 89. *JSTOR Journals*. Web. 11 Mar. 2016.

Wake, William C. "Sentence-Length Distributions of Greek Authors." Journal of the Royal Statistical Society. Series A (General) 1957: 331. JSTOR Journals. Web. 11 Mar. 2016.

Brocardo, Marcelo Luiz, et al. "Authorship verification for short messages using stylometry." Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on. IEEE, 2013.

Peter Millican. Signature Stylometric System 2003
http://www.philocomp.net/humanities/signature.htm

Evaluating Variation in Language Laboratory. Java Graphical Authorship Attribution Program (JGAAP) October 15, 2012 http://evllabs.com/jgaap/w/index.php/Main_Page

Eder et al. Stylo 2013 https://sites.google.com/site/computationalstylistics/stylo