

Rachel Hogue

Statistical Analysis and Comparison of Texts in Ancient Greek

The task of a classicist is similar to that of a detective: as classicists study ancient literature, their goal is to put together the pieces of history through their findings. They search for patterns or echoes in historical writings that may provide a clue that will help to place the text being studied into a broader historical context. They hope to discover some piece of the puzzle in the pages of literature that remain from antiquity that will allow them to come to a conjecture about the time period, the people that lived during that time, or about the artifacts that have endured to present day.

This project is inspired by Professor Wareh, a Classics professor at Union College. He was intrigued that the phrase “autonomous body” appears three times in all of Greek literature: twice in the works of the author Thucydides, once in those of Herodotus. As these two historians both lived in Greece during the fifth century BC, perhaps this discovery could be used as evidence that they influenced each other's writing. A person could certainly type “autonomous body” into a search engine of the Greek corpus to uncover this interesting fact, but how would he or she know to look for “autonomous body” in the first place? What if a tool existed that compared the two authors in order to obtain this result?

One of the inevitabilities of studying Greek and Roman authors is that scholars, particularly students who are just beginning and thus not yet experienced readers, become absorbed in and very concentrated on whichever particular body of work they are translating and/or studying. Placing a work or author into the whole of Greek literature or considering questions such as what authors may have influenced his/her writing, what authors did he/she influence, what words or phrases does he/she use that are typical or atypical of their genre or time period, is a difficult undertaking for a classicist. Thus, the broad goal of this project is to create a tool that will provide analytical access to texts through computational methods. Although inevitably some of the results will be insignificant connections that are merely coincidental, scholars can then resume an active role, investigating which results are

actually of consequence.

I will begin by presenting the existing resources available for Greek scholars. Next, I will state more explicitly the questions I wish to answer. After explaining the steps I have taken so far, the methods that I have used, I will review some testing results and discuss potential future work.

The Thesaurus Linguae Graecae (TLG), is a comprehensive digital library of ancient Greek literary texts. The library is available online via a subscription, as well as on CD-ROM. The library contains more than 105 million words. In comparison, the Wall Street Journal corpus, released 1987-89, contains 30 million words (<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43>). Today, however, corpora are substantially larger. For instance, the BLLIP North American News Text, another (newer) corpus of automatically parsed news texts from the 1990s, contains approximately 350 million words. The Google corpus contains over one trillion words. Thus, the TLG is moderately sized; however, the texts are split over 4,000 authors, so that the number of words per author is relatively small.

The TLG has been utilized by several different projects in the classics arena. The Perseus Project, created by Gregory Crane, Neel Smith, and Gabe Weaver, members of the Classics Department at Tufts University, uses tools to navigate a subset of the TLG texts. This online resource contains morphological data and dictionaries for Greek and Latin words. Diogenes is another free online resource that contains morphological data and dictionaries. Developed by Peter Heslin of the Classics Department at Durham University (UK), this tool can perform searches limited to specific authors and works by date, genre, location, author's gender. However, Diogenes is simply an interface to access and search the TLG; the ancient Greek texts must be purchased separately.

The Questions

Although the basic search tools for ancient Greek texts presented above are capable of finding a word or a phrase within a specified proximity, the goal of this project incorporates an additional aspect to such searching: comparison. Existing tools are useful when an individual knows what it is that they

are looking for in the ancient Greek corpus. This tool, on the other hand, will provide insights that the user would not likely discover without it. Instead of answering a question such as, “Where in ancient Greek literature does Achilles appear?”, the user can find answers to a comparison question, such as, “What textual tendencies does the work of Herodotus share with the works of the ancient Greek medical authors?”

I focused on developing computational techniques capable of solving two preliminary questions: (1) What words or phrases make an author unique in comparison to his/her peer group? and (2) What words or phrases make two authors *similar* in comparison to their peers? The first question attempts to find literary tendencies that are common to a single author and distinguish that author from his/her peer group. The second adds a further comparison factor, and attempts to find literary tendencies that are common to *two* authors, but are uncommon in the literature of their peer group.

Completed Tasks

The process consists of several major steps, which will each be explained in-depth: (1) text clean-up, (2) developing a technique to extract unique words and phrases for a single author, (3) variant ngram definitions, (4) developing a method to compare authors, and (5) testing.

Step 1: Text Clean-Up

Ancient Greek does not use the Latin alphabet; thus, the authors of the TLG had to develop a pattern for transcribing the ancient Greek to ASCII, in other words, a way to write Greek using the Latin alphabet. Unfortunately, much of this formatting is insignificant to the text analysis that I am creating methods for, and adds characters which only complicate the process of text manipulation. In addition, the TLG files contain symbols which specify formatting and references. Thus, the first step was cleaning up the text. Once this text editing was complete, I was able to focus on the next goal: finding words or phrases that are special for a particular author.

Step 2: TF-IDF

In order to find phrases that are special for a particular author, I used a method called TF-IDF.

The acronym TF-IDF stands for *term frequency, inverse document frequency*. The TF-IDF score for a sequence of words is, plainly, that sequence's frequency in a document multiplied by the inverse of that sequence's frequency in a background collection. There are some commonly used linguistics terms for these word sequences. A bigram consists of two consecutive words, say word A and word B. Word order matters; that is, “A B” is a different bigram than “B A”. Similarly, a trigram consists of three consecutive words, a fourgram of four consecutive words, and so on and so forth. Thus, an ngram consists of n consecutive words.

In TF-IDF, the term frequency for an ngram is the number of times an ngram is used by an author divided by the total number of ngrams in the author's writing. The first step in computing the term frequency is to compute ngram frequencies, which is simply counting ngrams for authors. For instance, Thucydides uses the bigram “autonomous body” two times. However, an ngram may be used more times by an author merely because the author wrote more. An example in English is the bigram “he said”. If author X writes a five page short story, and author Y writes a 500 page novel, author Y is probably going to use the phrase “he said” notably more often than author X. Thus, the ngram frequencies must be *normalized* so that they do not rely on length, and so that they are values that can reliably be compared. The calculation to normalize them is simple: each ngram frequency is divided by the total number of ngrams written by the author.

Unfortunately, this method presented a problem: the majority of the results included one very common word. Many of these were articles, prepositions, conjunctions, etc. In linguistics, these are referred to as “function words”. The inverse document frequency, then, is an added weight which takes into account how many authors actually use a particular ngram. Common words used by all authors are given very little importance.

Step 3: Variable Ngram Definitions

As mentioned previously, in linguistics, an ngram is a series of consecutive words. Two ngrams are the same if and only if they contain the same words in the same order. However, one of the most

interesting aspects of Greek is its variable word order. In English, the function of a word is designated by its position in a sentence. For instance, in the sentence, “Billy ate the apple,” we know that Billy is the subject because he appears first, before the verb, and that “the apple” is the direct object, because it follows the verb. If we said instead, “The apple ate Billy,” the functions of “the apple” and “Billy” would be reversed, and the meaning completely changed. The apple, instead of Billy, is now performing the action, and Billy is receiving the action. In Greek, however, the function of a word in a sentence does not depend on its position in the sentence; instead, its role is determined by its form. You could switch the position of two words, as we did in the English example, without changing their functions or the overall meaning of the sentence. Thus, I made the option of setting a more flexible definition of “ngram”. I first made the definition of ngram allow variable word position. When using this definition, ngram A-B is considered to be the same as bigram B-A.

Secondly, I adjusted the code to take into account the component words of an ngram within a broader context. For instance, in English, we may say, “the little, adorable puppy.” However, we could also say, “The puppy, who was little and adorable.” In this case, we would want to count the trigram “little, adorable puppy” twice, instead of having two separate trigrams (“little, adorable puppy” in the first example, and “The puppy, who” as well as “little and adorable” in the second) that are each counted once. Thus, when this definition of ngram is used, the code counts an occurrence of ngram X, consisting of words $\{n_0 n_1 \dots n_i\}$, if word n_k occurs y words after word n_{k-1} before any punctuation is reached. The spread, or window, is determined by the user, and this is the number of words within which an ngram can occur.

Step 4: Comparison

I have discussed methods that I implemented in order to determine what makes one author special. However, my second goal incorporates a second author. For this step, I developed a method using a graph illustration that determines in what way two authors are similar in comparison to their peers. Each point on the graph represented an ngram; the x-axis values are Author 1 TF-IDF scores,

while the y-axis values are Author 2 TF-IDF scores. The ngrams that are interesting have two qualifications: (1) they are close to the line $f(x) = y$, because these ngrams have the most similar frequencies for both authors, and (2) far away from the origin (0,0), because the TF-IDF scores must be high for both authors, meaning these are the most unique ngrams in comparison to the authors' peer group.

For this method, I used a reference point. This is a point that is on the line $f(x) = y$ and is the maximum TF-IDF score of both authors: $(x,y) = (\max(\text{auth1}, \text{auth2}), \max(\text{auth1}, \text{auth2}))$. The distance of each point, each ngram, from this reference point is computed, and the ngrams are then sorted such that the points that are closest to this reference point are at the top of the list. Users can then view this list, or they can specify a certain top percentage of them to graph.

Step 5: Testing

For the testing phase, I manipulated six variables: author1, author2, background corpora, stopword percentage threshold, variant order ngrams, the size of the ngram, and the window size of the ngram. As tests currently take upwards of fourteen hours, I minimized the tests that I performed using the following steps:

- (1) Choose two sets of authors and subcorpora to use.
- (2) Determine a stopword percentage threshold to use for all tests; run each test once using this stopword percentage, and once without using stopwords.
- (3) Work with bigrams and figure out what is the best window of words (e.g. 2..5 words) to look at if you are either ignoring word order or fixing the word order.
- (4) Repeat step 3 for trigrams

For step one, I chose to compare Demosthenes and Aeschines against the background subcorpus of Attic Prose and to compare Herodotus and Thucydides against the background subcorpus of Historians. Next, I ran tests to determine a stopword percentage threshold to use. Originally in computing stopwords, my code determined the IDF score of an ngram based on the particular peer

group an author was being compared to. For instance, for Demosthenes and Aeschines, the IDF score for an ngram was based on its occurrence in the fourteen Attic Prose author files. However, even in the stopword file generated with a threshold percentage of 100%, many words were present that would not typically be considered stopwords, but instead content words. This method only works really well, if we are looking at a significant number of documents; fourteen documents is not substantial. Thus, I instead calculated the list of stopwords using the whole 1,825 file collection.

Although results from these tests have not yet been analyzed, some preliminary tests have revealed promising results. For example, when comparing Herodotus to the Hippocratic Corpus, a collection of medical works associated with the Greek physician Hippocrates, words such as *disease*, *the body*, and *according to nature* were much more common in the Hippocratic corpus than in the writings of Herodotus; on the other hand, phrases such as *O King* (opening address of speeches), *to Delphi*, *the Greeks*, and *the Persians*, appeared much more frequently in Herodotus' work than in the Hippocratic Corpus. We also found in the top-ten non-trivial results the phrases "I can with certainty" [say / judge / determine] and "I came to". Both of these, and especially the latter, reflect something significant about how Herodotus presents himself as an authority in his own authorial voice. These results confirm the potential of the algorithms, with appropriate adjustments, to recognize thematically important phrases.

Future Work:

I have successfully implemented TF-IDF, which returns words and phrases which that make an author unique in comparison to his/her peer group. I have determined an effective way to incorporate a second author that quantifies the similarity of the words and phrases used by the two authors in comparison to their peers. However, there is much work that can still be done on this project.

One outstanding problem is the presence of variant word forms in the results. One of the primary objectives in the continuation of this project is to take different word forms into account. We have access to the Greek corpus parsed, from the Perseus Project. This resource gives lemma

information (person, number, tense, mood, voice, gender, case, degree). With this information, function words can be eliminated using these other available tools, and variant word forms can be associated with their root form.

The TF-IDF method can also be enhanced by considering specific documents rather than authors. The greater the number of documents used in creating the IDF corpus is, the more precise and meaningful the results should be. This entire project has been coded for authors. Splitting the TLG author files into documents would likely significantly improve results.

In addition to enhancing TF-IDF, we have recently become aware of a similar tool for Latin titled Tesseractae. Indeed, the website for this tool reads, “The Tesseractae project aims to provide a flexible and robust web interface for exploring intertextual parallels. In a basic search, selected works of Latin authors can be compared. Phrases from the texts which match in at least two of six relatively unfrequent words are grouped together for comparison, with links to their original context” (<http://tesseractae.caset.buffalo.edu/index.php>). Trying the Tesseractae web site is perhaps the quickest way (through Latin) for anyone with an interest in Classics to get an idea of the kinds of results somewhat statistical methods can suggest. Regarding SACTAG, transcribing the algorithms and code used by Tesseractae to be used with Ancient Greek would likely be very useful, as Ancient Greek and Latin are similar in structure and form.

Finally, a crucial step for this project is the design of a user interface. When this is complete, students will be able to utilize these scripts. This is a tool that anyone could use, and Professor Wareh is eager to pilot the tool to get students busy testing the hypotheses of significant connections it may generate.