Computational Stylometry Applied to Translations of Leo Tolstoy's *War and Peace*

AP Research

Word Count: 5243

30 April 2024

Introduction

In 1869, Leo Tolstoy published his masterpiece, *War and Peace*, in its original Russian. Yet, it would be another seventeen years before Clara Bell's original English translation appeared (Murphy). Clara Bell's version displayed flawed characteristics--itself having been an English translation of a French translation--and so a series of successive translations followed. Over the next century, a diverse cast of translators took on the task of converting *War and Peace* into English. Today, there exists a multitude of translations that are far evolved from Clara Bell's. Accompanying this abundance of choices is online debate about which translation readers should choose. Lucy Fuggle of *Tolstoy Therapy* personally claims that Anthony Briggs has the best, most engaging translation (Fuggle). Shilpi Agarwal of *ThinkSync* holds that the Aylmer and Louise Maude translation strikes the right balance of authenticity and language fluidity (Agarwal). And across Reddit, a jungle of opinions assert a contrasting superior translation ("Best English Translation of War and Peace?" Reddit). This debate suggests that the translations of *War and Peace* offer distinctive advantages and that their writing style and use of language differ from each other. In contrast, there exists a realm of separate debate arguing that translators exert minimal influence on their original work. The most pertinent example of this argument comes from the case of Lenita Esteves, who translated the popular novel, *The Lord of the Rings*, into Portuguese. Esteves translated the novel in the 1990s prior to the release of *The Lord of The Rings* movie which turned the book into a bestseller. Esteves was surprised to discover that the Brazilian version of the movie used key components of her translation without her consent. Therefore, Esteves sued both the publishing house and the movie distributor, claiming a right to her intellectual property. The publishing house contested her claim, asserting that translators are not paid for copyright, but simply for the task of translation itself (Esteves). Thus, two avenues

of how translations can be understood emerge. In the case of *War and Peace*, it would appear that translators create works inhabiting independent and unique literary styles, but in the example of Esteves, it seems that translators--and the literary merit of their translations--are viewed as dependent on an original work.

Initially, this dilemma appears to be based on subjective viewpoints and thereby unanswerable. However, in the field of computer science, a study known as stylometry has recently surfaced. Stylometry is the statistical analysis of an author's writing style via computational analysis of a text's key literary elements. Although stylometry has extensive use across a variety of disciplines, its application to translations is limited.

Unlike within other disciplines, there are additional challenges associated with translator stylometry because translators tend to hold limited freedom of expression while translating, opting to instead be as invisible as possible in their writing. Additionally, many prior studies have failed to identify translator stylometry. Mikhailov and Villikka's study proposed that computational methods can not properly identify a translator stylometry (Mikhailov and Villikka). Further, Hedegaard and Simonsen initially considered a translator's effect on a text to be noise that covered the original author's stylometry instead of a unique intellectual contribution (Hedegaard and Simonsen). Rybicki supports the findings of Mikhailov and Villikka when concluding, by clustering analysis, that translations were grouped based on their original authors rather than their translators (Rybicki). However, one study by El-Fiqi et al. asserts that a translator stylometry exists, but also maintains that their findings are merely an impetus for further research to support their conclusions, which are nearly isolated within surrounding literature (El-Fiqi). I hope to contribute to this debate by further supporting El-Fiqi et al.'s findings with a stylometric analysis of the translations of *War and Peace*, which will demonstrate

or refute the existence of translator writing styles computationally. Because *War and Peace* has a large English-translation corpus, it is an ideal candidate for translator stylometric analysis.

The central question for this research study therefore becomes: "using computational linguistics (stylometry), can a unique difference between the writing styles of the translations of *War and Peace* be identified?"

To answer this question, I employ a mixed-method stylometric analysis in which six *War and Peace* translations are divided into their first ten chapters. These first ten chapters of each translation are then divided further into chunks of fifty sentences. Each of these chunks is then run through various quantitative stylometric algorithms using the Python programming language. From the returned values of these stylometric algorithms, a Principal Component Analysis (PCA) graph is generated to analyze results and determine, by qualitative observation, if there is a uniqueness between the writing styles of the *War and Peace* translations or not.

Literature Review

Authorship Attribution

Stylometry is founded on the principle that everybody has a unique writing style, much like a fingerprint. To support this viewpoint, Kestemont defines Stylometry as the quantitative study of (literary) style (Kestemont). Stylometry is, therefore, a highly versatile tool, and looking at examples of its relevant past usage and success greatly reveals its capabilities.

Stylometry is often implemented for purposes of authorship attribution, the process of assigning a literary work (typically a disputed one) to an author through stylometric means. Although my research will not focus on authorship attribution, this review of literature will still

cover authorship attribution because it is a highly influential branch of stylometry, and looking into it provides valuable insight on stylometry's relevant background. Patrick Juola, a leading expert in stylometric research, references authorship attribution as such:

> Applications of authorship analysis include not only the literary, but also the historic, the journalistic, and even the legal. This technology has proven to be a useful scholarly tool across the academy (Juola).

Juola's words ring especially true for Mostellar and Wallace's landmark study on *The Federalist Papers*. Of the 85 essays collected in *The Federalist Papers*, 12 have disputed authorship primarily between Alexander Hamilton and James Madison. To assign authors to these disputed papers, Mostellar and Wallace developed a stylometric algorithm to identify key literary components of both Hamilton and Madison's writings and compare these identified components to the disputed papers. The study was successful, and Mostellar and Wallace were able to reasonably assign the disputed papers to the authors (Mostellar and Wallace).

Taking on a challenging task, Battles applied stylometry to the identification of the literary relationship between Old English and Old Saxon poems. The poems hold a convoluted and controversial history regarding the nature of their authorship and circulation, but by utilizing stylometric n-grams, Battles determined that Old English and Old Saxon poetry showed an extremely close relationship (Battles).

There are also instances of successful stylometric research beyond authorship attribution. In a recent study, Ríos-Toledo et al., researchers associated with Tecnológico Nacional de México and the Instituto Politécnico Nacional, concluded that tracking the frequency of n-grams is a reliable measure for determining an author's writing style, and that with n-grams, it is possible to detect changes in writing style over time. The study

concluded that noticeable distinctions in writing style can be identified along the span of an author's career (Ríos-Toledo et al).

Still, although a large body of successful research exists regarding stylometry overall, there is a distinct lack of research regarding its specific application toward translators. Considering the focus of this study, it is doubly important to look into previous research addressing translator stylometry.

The Challenge of Translator Stylomery

Multiple studies have tackled the challenge of translator stylometric analysis. However, many prior studies are either unsuccessful or are limited, which invites further research, such as my own, to expand onto the growing body of translator stylometry.

In a 2001 study, Mikhailov and Villikka examined the existence of translator stylometry. The research based itself on a corpus of Russian works translated into Finnish. Utilizing vocabularic richness, word frequencies, and favored words, the pair demonstrated that the language of the translation formed a closer relationship with the translation than did the translator. The study concluded that although there are stylistic features relevant to translators, they were not able to prove the existence of translator styles (Mikhailov and Villikka). Therefore, my study utilizes a new stylometric method to establish a separate avenue of stylometric analysis and results.

In 2012, Rybicki stylometrically analyzed a variety of English translators of foreign works and foreign language translations of English works. Rybicki's stylometric analysis centered around using Burrow's Delta transformed into Cluster Analysis tree diagrams. From his results, Rybicki concluded that his stylometric analysis gave little credence to the notion of a

translator's stylometric identity. To use the study's own words, "[n]ow this study seems to be adding an additional dimension to 'the translator's shadowy existence…" (Rybicki, page 14) Of course, however, all stylometric research rests on the method under which it is conducted, thus my study employs a different approach to draw out new results.

Previous unsuccessful attempts at identifying a translator stylometry illustrate an unpromising foundation, but there are still instances of stylometry applied to translators generating significant findings.

El-Fiqi et al, in their study "Network Motifs for Translator Stylometry Identification," demonstrate that through employing novel methods, it is possible to identify a translator's writing style. The researchers, understanding the limitations of prior studies, primarily utilized network motifs to detect the literary elements of translators. Ultimately, their research was successful, and El-Fiqi et al claimed that a translator's writing style could be identified through computational stylometry. However, room for further research remained. El-Fiqi et al themselves state that the first contribution of their research was to provide "evidence for the existence of translator stylometry" and that analysis could be extended to "a larger number of books and translators." (El-Fiqi, page 29). My study aims to do just that: extend the literature surrounding the computational stylometry of translators by analysis of *War and Peace*'s translation corpus.

However, all research regarding stylometry requires a robust methodology. Many methods have been tested against the challenge that is translator stylometry, and of those, many have failed to produce significant results. Thus, it is important to continually try new forms of methodology, which serve to advance how translator stylometry is applied. My study therefore implements a novel method that has yet to be tested against translator stylometry.

A New Approach

A study conducted by Elahiand and Muneer, researchers associated with FAST National University of Computer and Emerging Science, demonstrated that through analyzing three main categories of a text's literary features---those being lexical, vocabulary, and readability--multiple writing styles within a single work could be identified (Elahiand and Muneer). Displaying and examining results through a Principal Component Analysis (PCA) graph greatly aided in drawing conclusions. The study's focus on thorough scrutiny of a text's stylometry via a variety of analyzed textual features and the implementation of a PCA graph, which is designed to depict small, minute differences between points of data made Elahiand and Muneer's study a promising one. Therefore, the methodology used within the study is the model for my own research.

For the expanding field of translator stylometry, it is essential to address both new translations and methodology. That is, one must not only analyze unstudied literature but also employ different methods of doing so, which serve to establish further avenues for how translator stylometry can be understood. Thus, my study is directed at these two objectives. The translations of *War and Peace* act as unstudied literature, while a method derived from Elahiand and Muneer's study on the identification of writing styles, which, despite its promise, has yet to be used for translator stylometry.

Methodology

To investigate the proposed hypothesis, this research study, using the Python programming language, employed various formulas and techniques to perform a stylometric

analysis of selected English translations of *War and Peace* (for an example of the Python code written, see **Appendix Q**). Python was specifically chosen because of its statistical capabilities that conform to this study's purposes well.

The first component of this research was to select the English translations of *War and Peace* from which to derive literary components. The feasibility of this project relied on choosing which were suitable for analysis. I looked at three primary criteria to determine the suitability of a given translation, which provide an appropriate stage for the later stylometric analysis and are as follows: (1) The translator's relevance to modern-day readers of *War and Peace* (essentially, would somebody today reasonably consider reading the translation?) (2) The ease of access to online (e-book) versions of the translator's work, which are necessary to conduct a computer-based stylometric analysis within practical time. And, (3) The translator's credibility, which is determined by their association with notable publishing firms or organizations, or otherwise the translation's approval from Tolstoy himself. From these criteria, I selected the following translations as appropriate for the research inquiry, shown in **Figure 1** below.

Translator	Year of Publication
Nathan Haskell Dole	1889
Constance Garnett	1904
Alymer and Louise Maude	1922
Ann Dunnigan	1968
Anthony Briggs	2005
Richard Peaver and Larissa Volokhonsky	2007

Figure 1. Collected *War and Peace* Translations

Setup

Prior to applying the stylometric tests, all selected translations of *War and Peace* were divided into their first ten chapters. These first ten chapters were then further divided into chunks of fifty sentences. Each of these chunks was then analyzed using the stylometric tests, and values derived from the tests were returned (this portion acts as the quantitative aspect of the method.) From these returned values, Principal Component Analysis (PCA) graphs were generated to graphically represent the data and reach a conclusion through qualitative observations. Therefore, this study's methodology has three primary parts: dividing the *War and Peace* into their first ten chapters and chunks of fifty sentences, analyzing the chunks of fifty sentences with stylometric tests, and then using the values returned from the stylometric tests to generate a PCA graph of the data for analysis.

Stylometric Tests

Once the first ten chapters of each translation were divided into chunks of fifty sentences, the Python algorithm applied various stylometric functions to said chunks. For this study, the stylometric functions were obtained from Elahiand and Muneer's paper, *Identifying Different Writing Styles in a Document Intrinsically Using Stylometric Analysis* (Elahiand and Muneer). Many tests were also further supported by other relevant literature, especially if they were extracted directly from other research. Therefore, the following techniques are implemented into the Python algorithm employed within this research study. The techniques are divided into three categories of features: lexical, vocabularic, and readability, which all together form a coherent image of a work's stylometric identity.

Lexical Features

Average Word Length

The average word length function reports the total number of characters within each word in a text, divided by the total number of words. This returns a value of how complex a text's words are.

Average Sentence Length by Word

The average sentence length by word function reports the total number of words within each sentence in a text, divided by them by the total number of sentences. This returns a value of how complex a text's sentences are.

Average Sentence Length by Character

The average sentence length by character function reports the total number of characters within each sentence in a text, divided by them by the total number of sentences. This returns a value of how complex a text's sentences are.

Average Syllable per Word

The average syllable per word function reports the total number of syllables in each word in a text, divided by the total number of words. This returns a value of how complex a text's words are.

Functional Words Count

The functional words count function reports the total number of functional words in a text, divided by the total number of words. The frequency of function words in a text indicates an author's unique writing style.

Punctuation Count

The punctuation count function reports the total number of punctuation characters in a text, divided by the total number of words. This returns a value of how complex a text's content is.

Vocabularic Features

Hapax Legomenon

The Hapax Legomenon function reports every instance of a word that only appears once throughout a provided text, which provides a measure of a text's lexical richness: the more unique words there are, the more diverse a text's vocabulary is.

Hapax Dislegomenon

The Hapax Dislegomenon function reports every instance of a word that appears twice throughout a provided text, essentially, the opposite of the Hapax Legomenon function. The Hapax Dislegomenon therefore provides an additional measure of a text's lexical richness: the more repeated words there are, the less diverse a text's vocabulary is.

Type Token Ratio

The Type Token Ratio (Biber D.) function calculates the total number of unique words (vocabulary), divided by the total number of words within the text. Although simple, the Type Token Ratio test helps to measure how complex a text's vocabulary is.

Honores R Measure (Honores Statistic)

The Honores R Measure (Honores Statistic) (Honore A.) function records the number of words produced only once within a text (unique words), and returns a value representative of lexical richness based on the enumerated number of unique words. Honores R Measure generates its measure according to the formula, $R = 100 * \log(N / (1 - V_1/V))$, where V_1 is the number of unique words, V is the total vocabulary used, and N is the total text length. Higher values of R correspond to higher lexical richness.

Sichel's Measure

The Sichel's Measure (Sichel) function reports the number of words that appear twice throughout a text, divided by the total number of unique words (vocabulary), which provides a measure of a text's lexical complexity. The higher the value of Sichel's Measure, the less lexically complex a text is.

Brunet's Measure W (Brunet's Index)

The Brunet's Measure W (Brunet's Index) (Brunet V.) function reports a value of lexical richness according to the formula, $W = N^{V^{(-0.165)}}$, where N is the total text length and V is the total vocabulary used. Lower values of W correspond to higher lexical richness.

Yule's Characteristic K

The Yule's Characteristic K (Yule) function computes a value of lexical richness according to the formula $K = C[-\frac{1}{N} + \sum_{m=1}^{m_{max}} V(m,N)(\frac{m}{N})^2]$, in which *N* is the total number of words in a text, *V*(*N*) is the largest number of unique words, *V*(*m*,*N*) is the number of words appearing *m* times in a text, and m_{max} is the largest frequency of a word, anc C is a constant defined by Yule as $C = 10^4$. The larger Yule's K is, the less rich the vocabulary is (Kumiko).

Shannon Entropy

The Shannon Entropy function produces a measure of uncertainty of a probability distribution according to the formula $-\Sigma p_i log_2 p_i$, in which p_i represents the possible number of outcomes following the iterative, *i*. Although not directly related to stylometry, Shannon Entropy operates by determining the number of bits necessary to encode an *n* number of possible outcomes. Thus, if all vocabulary used within a text is understood as all possible outcomes, then a higher Shannon Entropy indicates a higher lexical richness.

Simpson's Index

The Simpson's Index function measures a given population's diversity according to the formula $D = \frac{N(N-1)}{\Sigma n(n-1)}$, when in the context of a stylometric study, *N* represents a text's total number of words, and *n* represents the total number of times a certain individual word appears within the text. Higher values of Simpson's Index, *D*, correspond to a higher lexical richness.

Readability Features

Flesch Reading Ease

The Flesch Reading Ease function quantifiably measures how difficult a text is to read according to the formula, [206.835 - 1.015*(total words / total sentences) - 84.6*(total syllables / total words)]. For the purposes of this research study, the Flesch Reading Ease test provides a coherent method of determining a text's readability.

Flesch-Kincaid Grade Level

The Flesch-Kincaid Grade Level function estimates the grade-level of a written work according to the formula, [0.39*(total words / total sentences) + 11.8*(total syllables/total words) - 15.59]. Within the context of this research study, the Flesch-Kincaid Grade Level test serves to define a text's readability.

Gunning Fog Index

The Gunning Fog Index function returns a value between 6 and 17 that corresponds to a text's readability according to the formula, 0.4*[(words / sentences + 100*(complex words / words)], with complex words defined as words consisting of three or more syllables. The Gunning Fog Index produces a clear analysis of a text's reliability and is therefore useful for the purposes of this research study.

Dale Chall Readability

The Dale Chall Readability Function returns a value between 4.9 and 9.9 that represents how easily a text can be comprehended according to the formula, [0.1579*((difficult words /

words) * 100) + 0.0496*(words/sentences)], with difficult words defined as any words not on the established word list, and should the percentage of difficult words be above 5%, then 3.6365 is aggregated to the returned score. The Dale Chall Readability function is useful for this research study for its capability to quantify a text's readability.

Principal Component Analysis

To analyze and draw conclusions from the generated data, a Principal Component Analysis (PCA) graph is created by the algorithm. A Principal Component Analysis graph works by smudging higher dimensional vectors into more easily-viewable dimensions, such as 3D or 2D. For the research study, there are twenty stylometric tests, which therefore form a twenty-dimensional vector. A PCA graph will take this twenty-dimensional vector and convert it to two dimensions, which enables the data to be visualized. Once a PCA creates visualized data, its analysis is based on distinct clusters of points that indicate a difference between variables, which in this case, the variables are the *War and Peace* translations. Shown below in **Figure 2** is an example of a PCA graph, which was formed in Elahiand and Muneer's study. The graph displays distinct clusters of points, therefore implying that there are two distinct writing styles. Within my study, a PCA graph is utilized to distinguish the writing styles of the translations of *War and Peace*. If distinct clusters form, then, by qualitative analysis, it can be concluded that the translations of *War and Peace* demonstrate unique writing styles from each other.

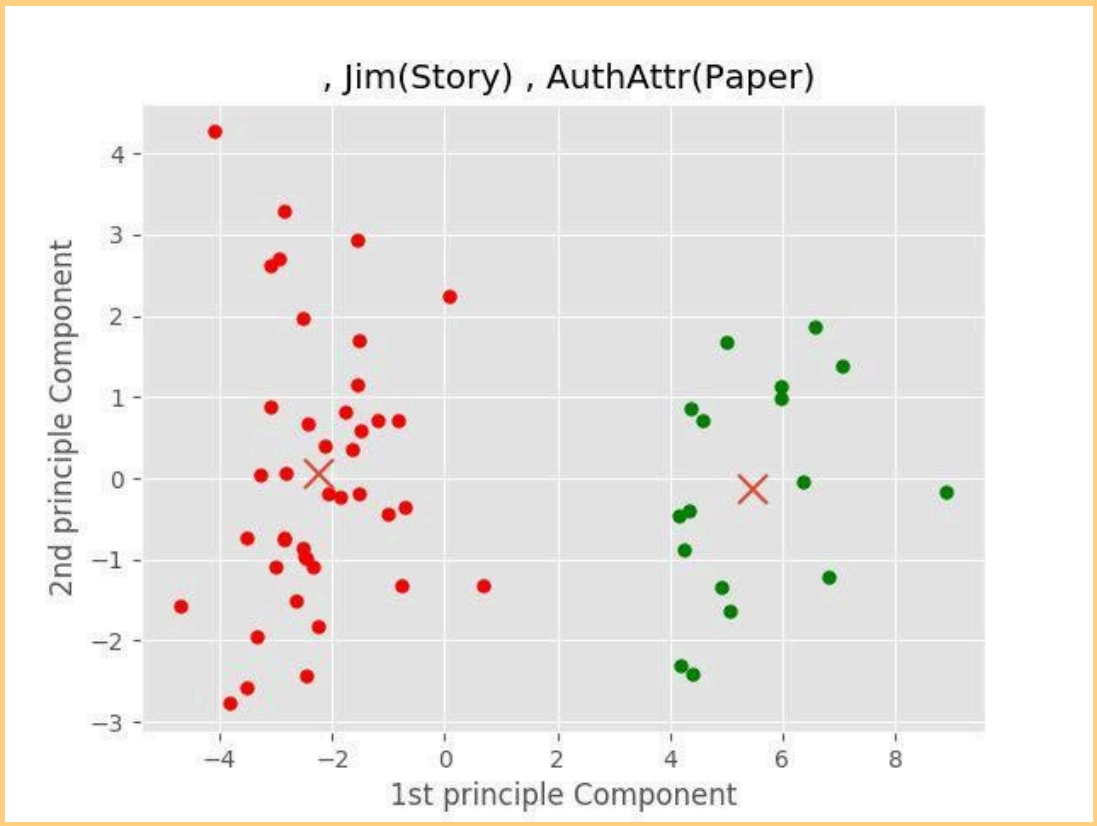


Figure 2. Elahiand and Muneer's PCA graph from their research study, *Identifying Different Writing Styles in a Document Intrinsically Using Stylometric Analysis*.

Results

As covered in the methodology section, the selected *War and Peace* translations were each divided into chunks of fifty sentences and then analyzed using twenty stylometric tests. Data generated from this procedure is displayed, by translation, as a Principal Component Analysis (PCA) graph, which maps out every chunk by smudging the twenty stylometric tests applied to them into two dimensions. From this procedure, the following PCA graph, shown in **Figure 3** below, was created.

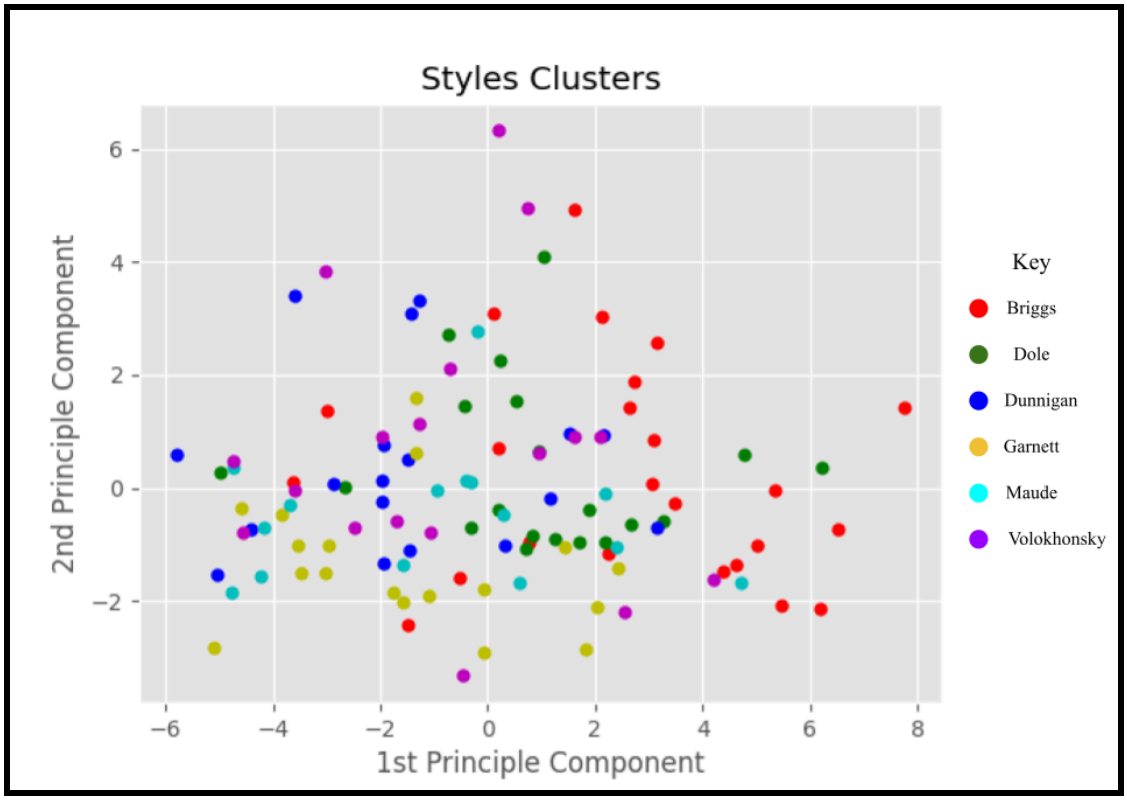


Figure 3: Overall PCA Graph of *War and Peace* Translations

The PCA graph displays no distinct clusters of points, which implies a lack of a distinction between the translations of *War and Peace*. However, this setback did not deter my research. By their nature, PCA graphs are often subject to information loss, thus a closer analysis of each translation compared to a single other translation was conducted. A PCA graph of this procedure is shown in **Figure 4** below.

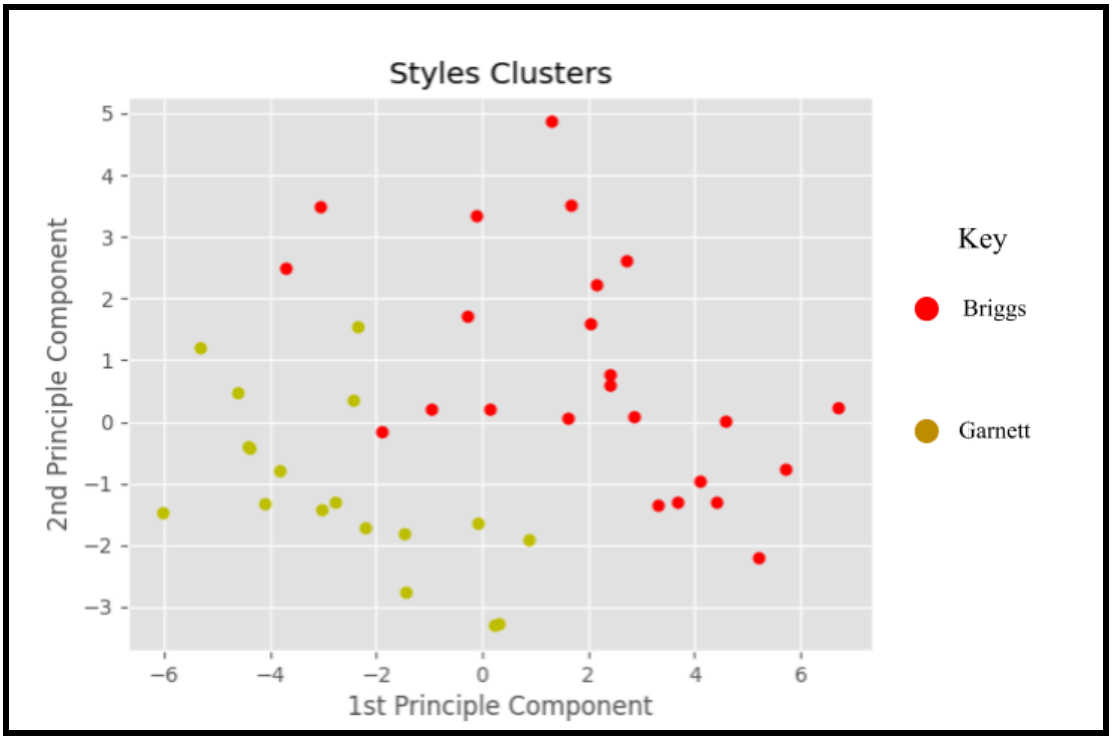


Figure 4: PCA Graph of the Briggs and Garnett Translations

Figure 4, which compares only the Briggs translation to the Garnett translation displays far more identifiable clusters of points. These identifiable clusters imply the opposite of what was established earlier within **Figure 3**. However, the results generated within **Figure 4** take precedence of those of **Figure 3** because they are derived from a more precise analysis of a comparison between only two translations, instead of every translation against each other at once. To further the findings produced by **Figure 4**, all translations studied were run against each other individually, as shown in the **Appendices**. Overall, when the translations are analyzed through individual comparison against each other, distinct groupings of points are thoroughly identifiable. Thus, it is clear through PCA graphing that there exists a unique difference between the writing styles of the translations of *War and Peace*.

Some generated PCA graphs make the difference in writing style between translations glaring. Shown in **Figure 5** below is the Dole translation run against the Volokhonsky translation.

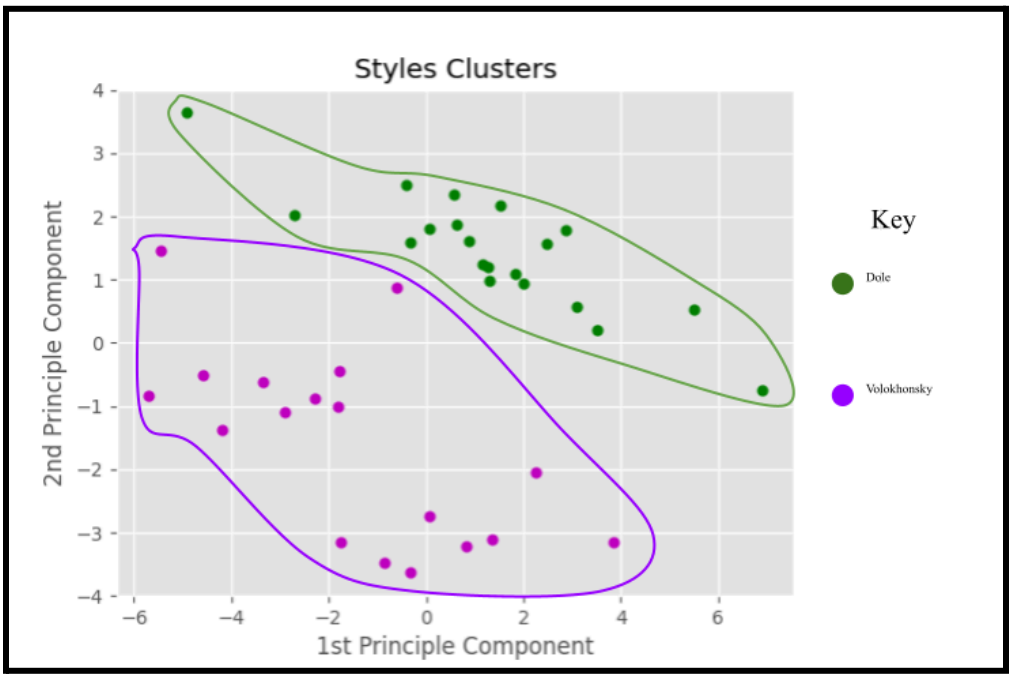


Figure 5: PCA Graph of the Dole and Volokhonsky Translation

In instances such as **Figure 5**, the separate clusters that define PCA graph analysis are made clear and are accordingly highlighted within the graph. Reviewing the generated PCA graphs, many display features similar to that of **Figure 5** (see Appendix B, D, H, I, J, K, and O), which make the distinctions between the writing styles palpable. While other PCA graphs did not establish as highlighted of a division, noticeable clusters of points are still identifiable (see Appendix C, E, F, G, L, and N.) Therefore, among the PCA graphs derived from the *War and Peace* translations, identifiable differences in their writing styles are noticeable through qualitative analysis of distinct clusters of points, the basis for drawing conclusions from PCA graphs.

Examples like the Dole and Volokhonsky translations demonstrate a significant difference between writing styles, but still, others did not. Shown below in **Figure 6** below is a PCA graph of the Dunnigan translation run against the Volokhonsky translation.

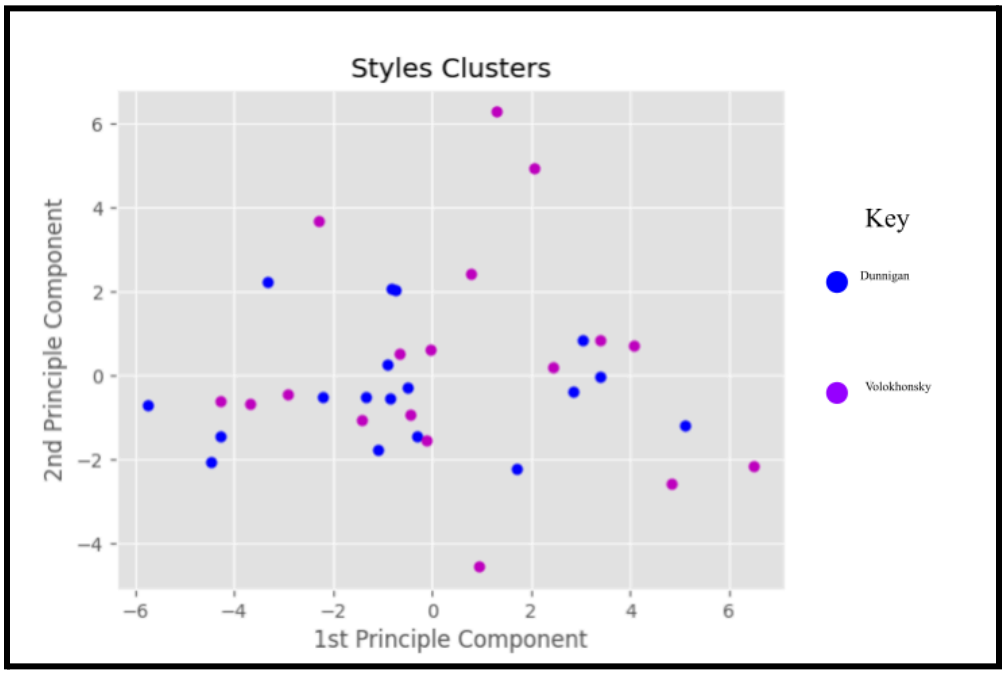


Figure 6: PCA Graph of the Dunnigan and Volokhonsky Translations

The Dunnigan translation shows significant overlap with that of Volokhonsky, which as established earlier in **Figure 3**, may serve as a reasonable argument against the initial hypothesis. However, even though the points display overlap, this does not necessarily disprove the initial hypothesis that the writing styles of *War and Peace* translations are separate from each other. Some overlap is to be expected within PCA graphs, which are not a perfect representation of high-dimensional data. On the same line of reasoning, PCA graphs are subject to a degree of information loss, which may slightly impact how results are interpreted. Thus, even though the PCA graph of **Figure 6** illustrates a lack of connection between the writing styles of translations,

its results must be understood within the context of the other PCA graphs, which do demonstrate noticeable differences, and the limitations that follow PCA graph analysis. Additionally, the points within **Figure 6** are not completely homogenous, in that an utter overlap between the two analyzed translations is not present, which implies that, at the very least, the writing styles of the translations are not exactly the same.

Figure 6 serves as an example of this study's limitations. The conclusions of this study are drawn from PCA graph analysis, which is subject to information loss, a factor that can skew findings. In addition, because PCA graphs rely on qualitative observations, especially in this case, it is difficult to analyze the statistical significance of their results. Although this does not necessarily play as a problem within my study because noticeable delineations within the PCA graphs can be identified and conclusions established from there, it is a component of my study's limitations that could be addressed within future research.

Although this study uses PCA graphing as its primary form of gathering conclusions, a table of averaged stylometric values was additionally generated to further support findings. The table displays the average mean values of each stylometric test applied to the chunks of fifty sentences that divided up the first ten chapters of each studied translation. The table, shown in **Figure 7** below, is less jarring than its PCA counterparts but does support the findings. When looking at the table, small discrepancies between the values of the stylometric tests applied to the translations appear, which support the notion that there exists a difference between the writing styles of the translations of *War and Peac*e. Little else was done with the table in **Figure 7**, however, because it is not the primary focus of my research's findings.

Translation	Briggs	Dole	Dunnigan	Garnett	Maude	Volokhonsky
Average Word Length	5.023	5.328	5.316	5.251	5.250	5.092
Average Sentence Length By Chunk	86.953	98.630	114.284	115.609	112.553	112.066
Average Sentence Length by Word	15.659	18.041	20.096	20.562	20.047	19.648
Average Syllable Per Word	1.602	1.663	1.693	1.684	1.660	1.623
Average Special Characters	0.001	0.004	0.003	0.003	0.002	0.002
Average Punctuation	0.029	0.039	0.027	0.028	0.026	0.029
Functional Words Count	0.577	0.620	0.617	0.604	0.608	0.596
Type Token Ratio	0.397	0.391	0.384	0.366	0.380	0.386
Hapax Legemena	0.309	0.310	0.302	0.280	0.291	0.304
Honore Measure R	673.302	682.553	698.122	700.796	696.518	697.266
Hapax Dislegemena	0.062	0.061	0.058	0.056	0.055	0.059
Sichele's Measure S	0.139	0.137	0.134	0.137	0.131	0.136
Yules Characteristic K	1355.020	1347.370	1271.540	1161.160	1213.770	1302.200
Simpson's Index	0.990	0.990	0.992	0.991	0.992	0.992
Brunet's Measure W	56.393	60.469	66.901	65.213	65.026	67.437
Shannon Entropy	9.053	9.118	9.184	9.060	9.087	9.272
Flesch Reading Ease	74.385	70.406	65.494	65.892	67.222	67.608
Flesch-Kincaid Grade Level	7.154	8.083	9.528	9.608	9.242	9.194
Dale Chall Readability Formula	10.115	9.704	10.438	10.238	10.297	10.960
Gunning-Fog Index	10.100	11.009	12.581	12.792	12.144	12.279

Figure 7: Table of Averaged Stylometric Features for each *War and Peace* Translation

In all, the results indicate that the inquiry posed at the start of this research, “using computational linguistics (stylometry), can a unique difference between the writing styles of the translations of *War and Peace* be identified?” is demonstrated to be true. Through multiple PCA graph analyses, the *War and Peace* translations form distinct clusters of points, which suggest that the translations have distinctive writing styles from each other, despite limitations imposed by a few generated PCA graphs. To further support this conclusion, a table of averaged stylometric values was created. Within the table, small discrepancies between the stylometric values of the translations emerge, thereby supporting that the translations are not of the same writing style. Thus, through my research, it can be established that translations of *War and Peace* demonstrate a difference in writing style from each other through a mixed-method stylometric analysis.

Discussion and Conclusion

A translator stylometry is identifiable between the translations of *War and Peace*. This conclusion is drawn from identifiable clusters within PCA graphs generated based on twenty stylometric components that were applied to the studied translations. To further support these results, a table that displays averaged values of stylometric components was created. Overall, these results support the conclusion that translations of *War and Peace* have unique writing styles from each other. However, this study's conclusions are not without their limitations. Due to processing capability, only the first ten chapters of each *War and Peace* translation were inputted. Should a more powerful computational device be available, the depth of the results could be expanded upon. Additionally, although PCA graphs provide a solid basis for investigating findings, they can be subject to information loss, which may skew results. To avoid this, my study applied a far closer analysis of translations examined on a one-to-one scale, but information loss within PCA graphs should still not be overlooked. Lastly, not all translations of *War and Peace* were studied because of constraints regarding their accessibility, for example, many paper-only translations were excluded based on their incompatibility with computational input, and so e-book variants were used to avoid this issue. An expansion of included *War and Peace* translations may have added to the findings produced by the study.

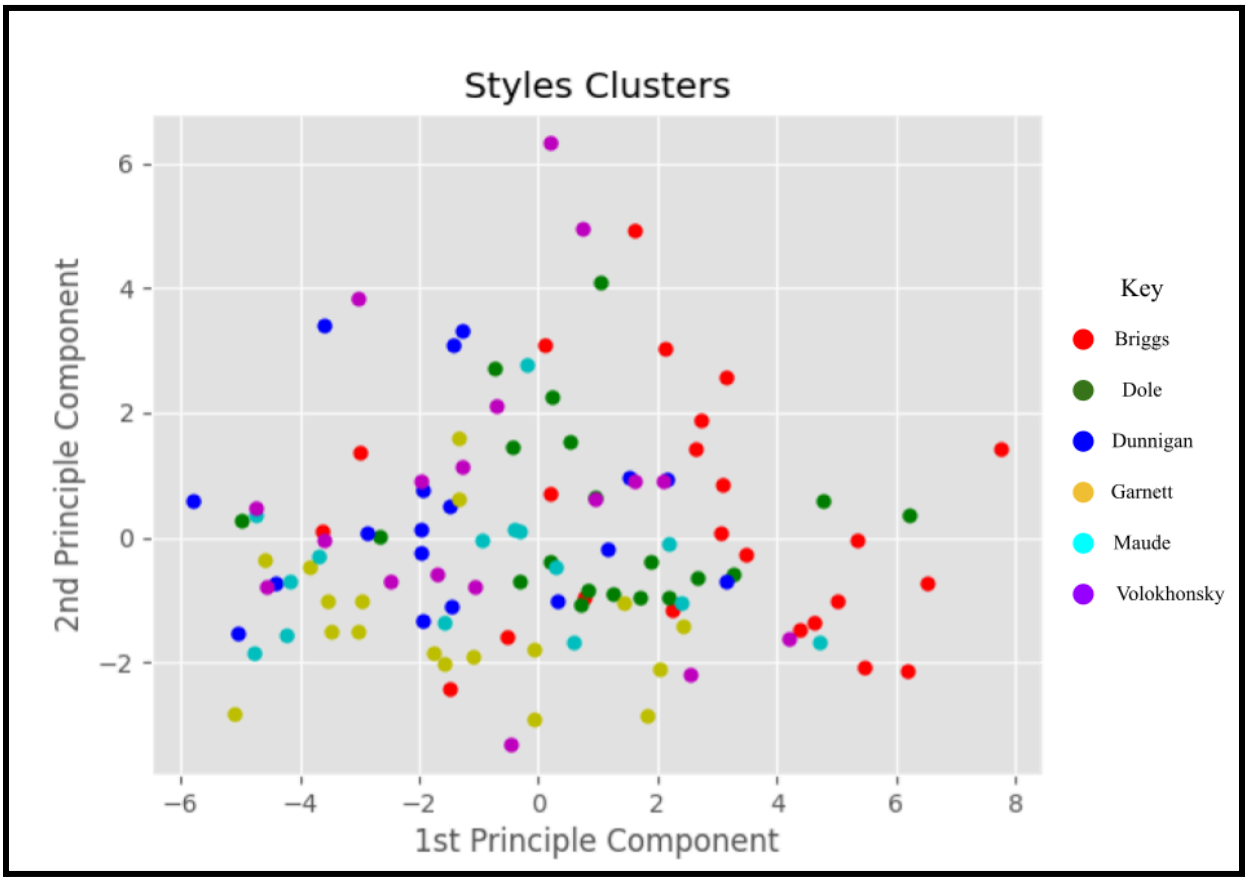
Despite the limitations, this study's conclusion resolved the initial inquiry that motivated my research, which was "using computational linguistics (stylometry), can a unique difference between the writing styles of the translations of *War and Peace* be identified?" Yet, further avenues for future study and implications of the results remain. As addressed in the introduction, the choice of analyzing *War and Peace* was partly spurred by the online debate surrounding differences in the writing styles of the translations. The conclusions of this study justify that one may identify individual qualities in the writing styles, but do not address what makes a

translation favorable or what the margin of dissimilarity between the translations are. Therefore, these unanswered inquiries can become the grounds for future research, which emphasizes building on top of the existing literature of translator stylometry instead of only expanding its breadth. In terms of implications, this study justifies a distinction in writing styles between translated work, which supports the intellectual property rights of translators, and aids those like Esteves (mentioned in the introduction) hold claim to their creative works as translators. Furthermore, by implementing a new method for analyzing translator stylometry, it becomes more viable to do so, providing a solid foundation for how to interpret and understand translator stylometry.

Alone, this paper contributes to providing evidence that translator stylometry exists and provides a valid methodology for conducting research relevant to translator stylometry. These factors provide an appropriate launching point for further research, and an expanded analysis of varying literature is encouraged.

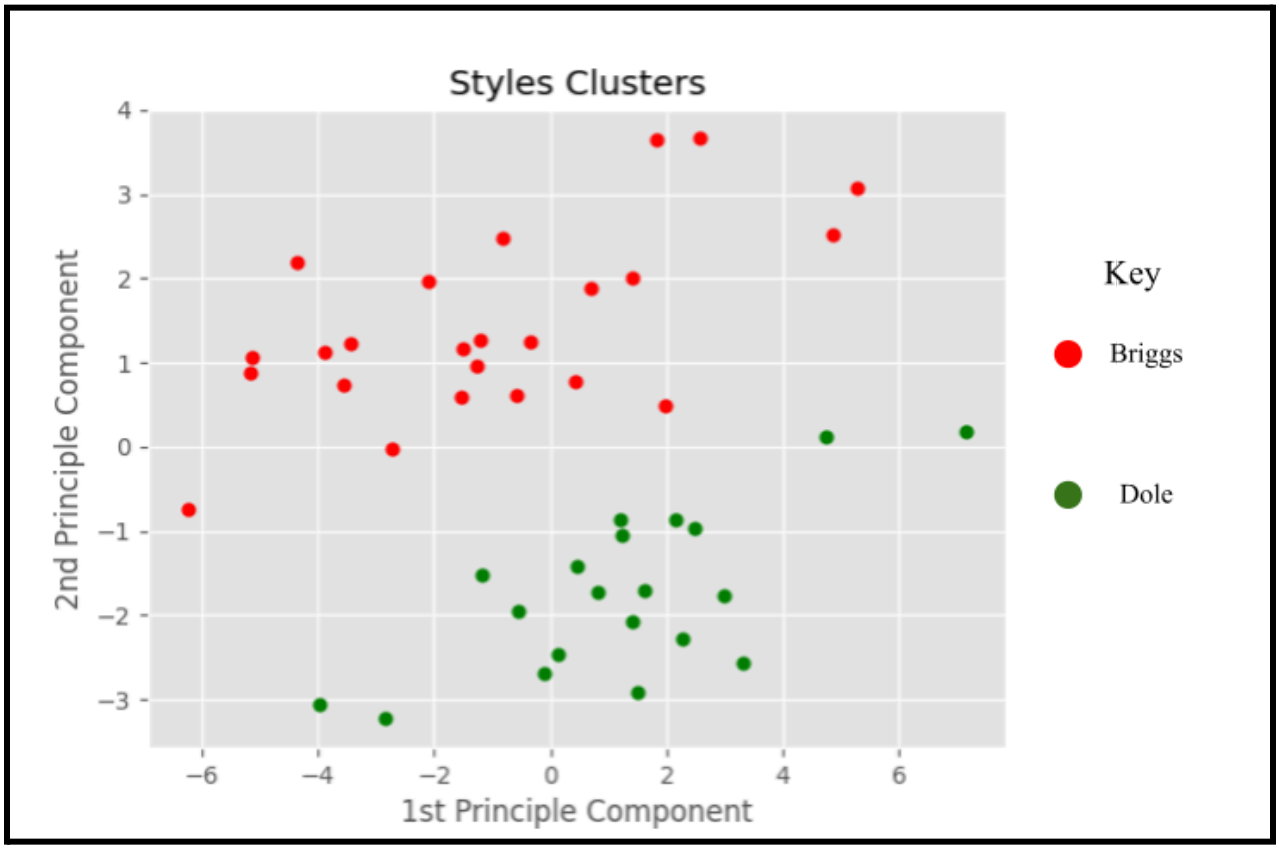
Appendix A.

Overall PCA graph of *War and Peace* translations (all translations run through stylometric tests against each other at once.)



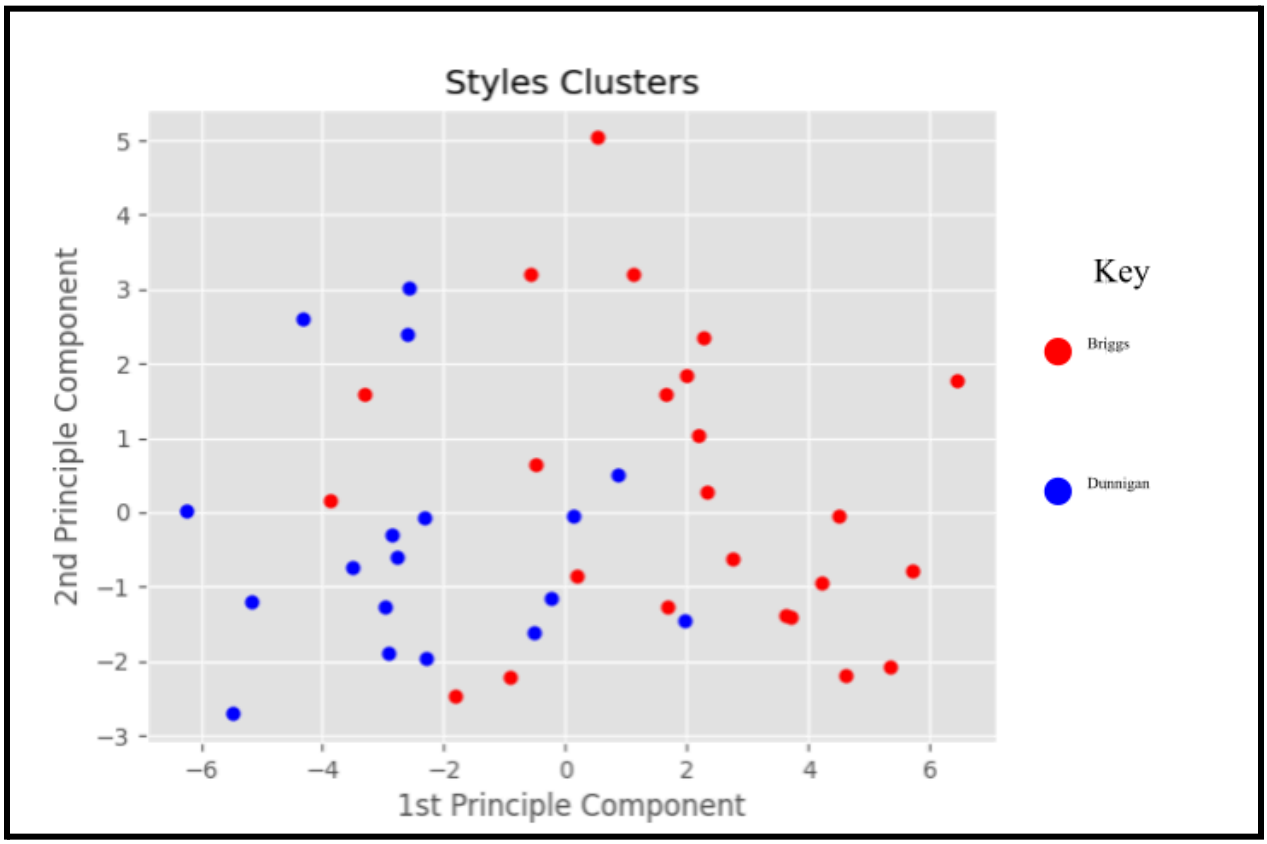
Appendix B.

PCA Graph of the Briggs and Dole Translations



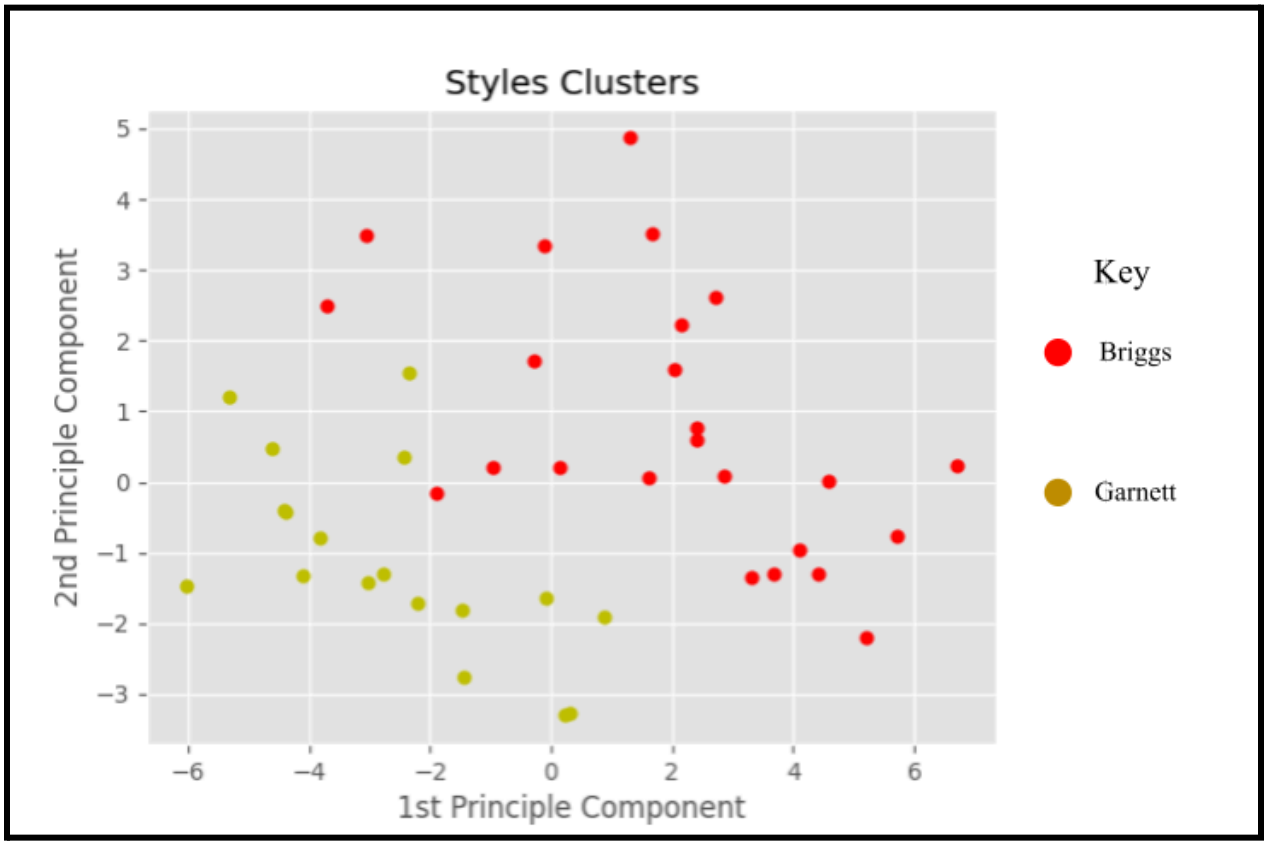
Appendix C.

PCA Graph of the Briggs and Dunnigan Translations



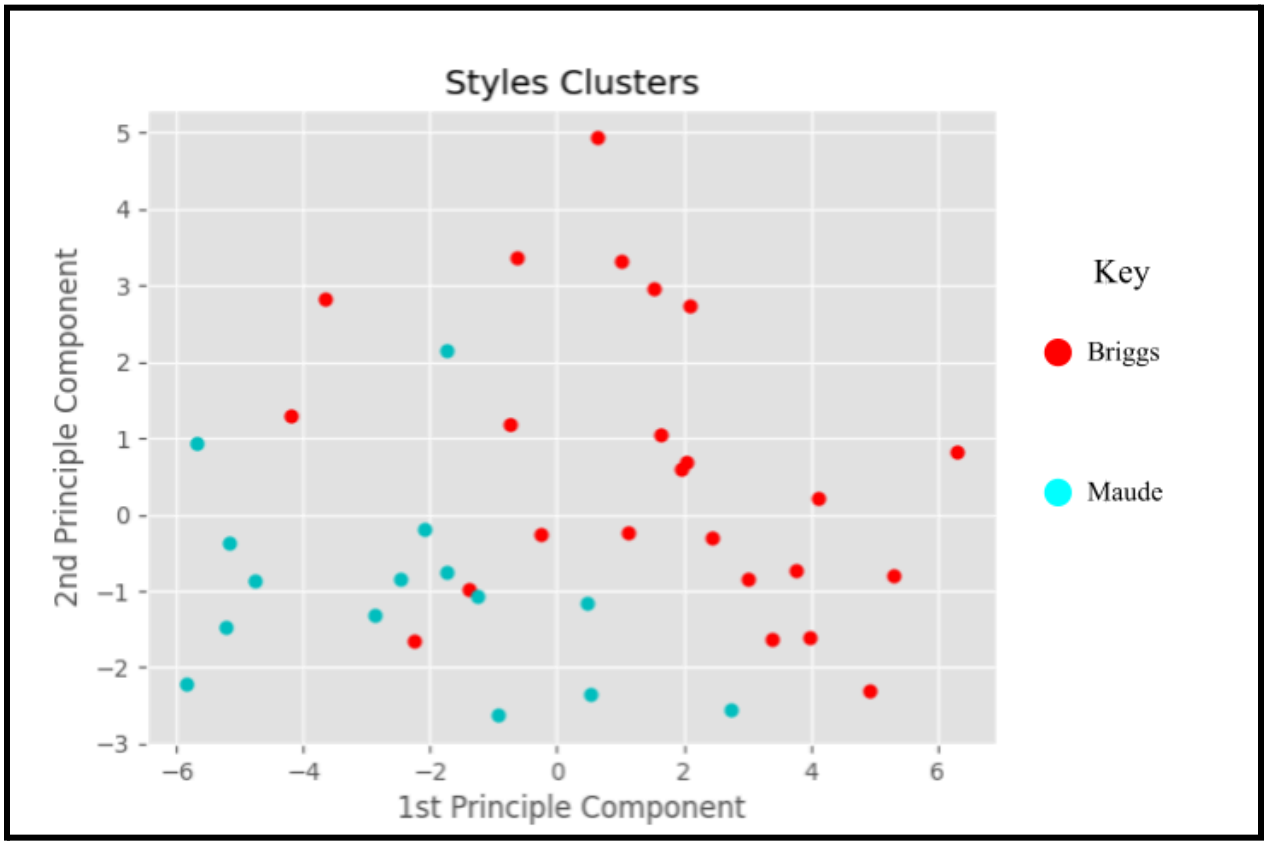
Appendix D.

PCA Graph of the Briggs and Garnett Translations



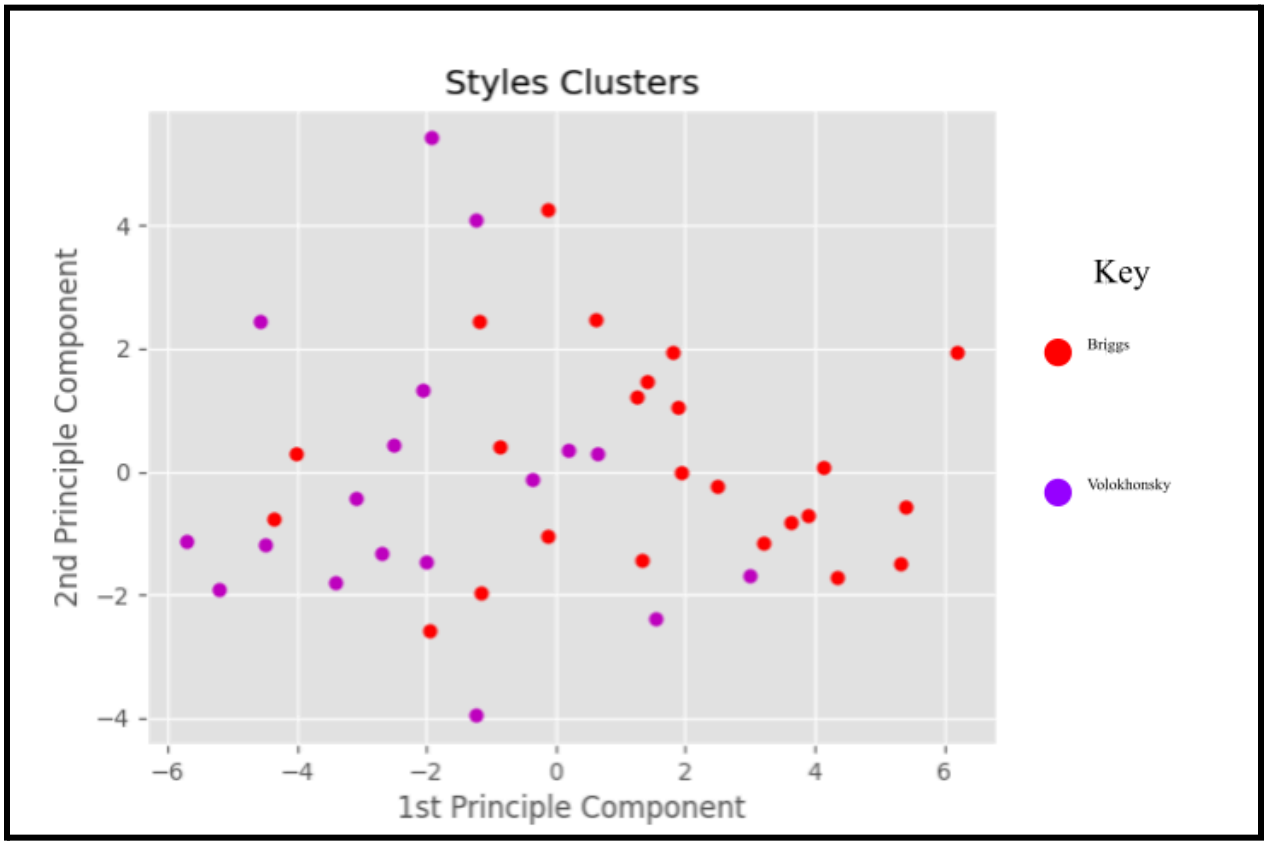
Appendix E.

PCA Graph of the Briggs and Maude Translations



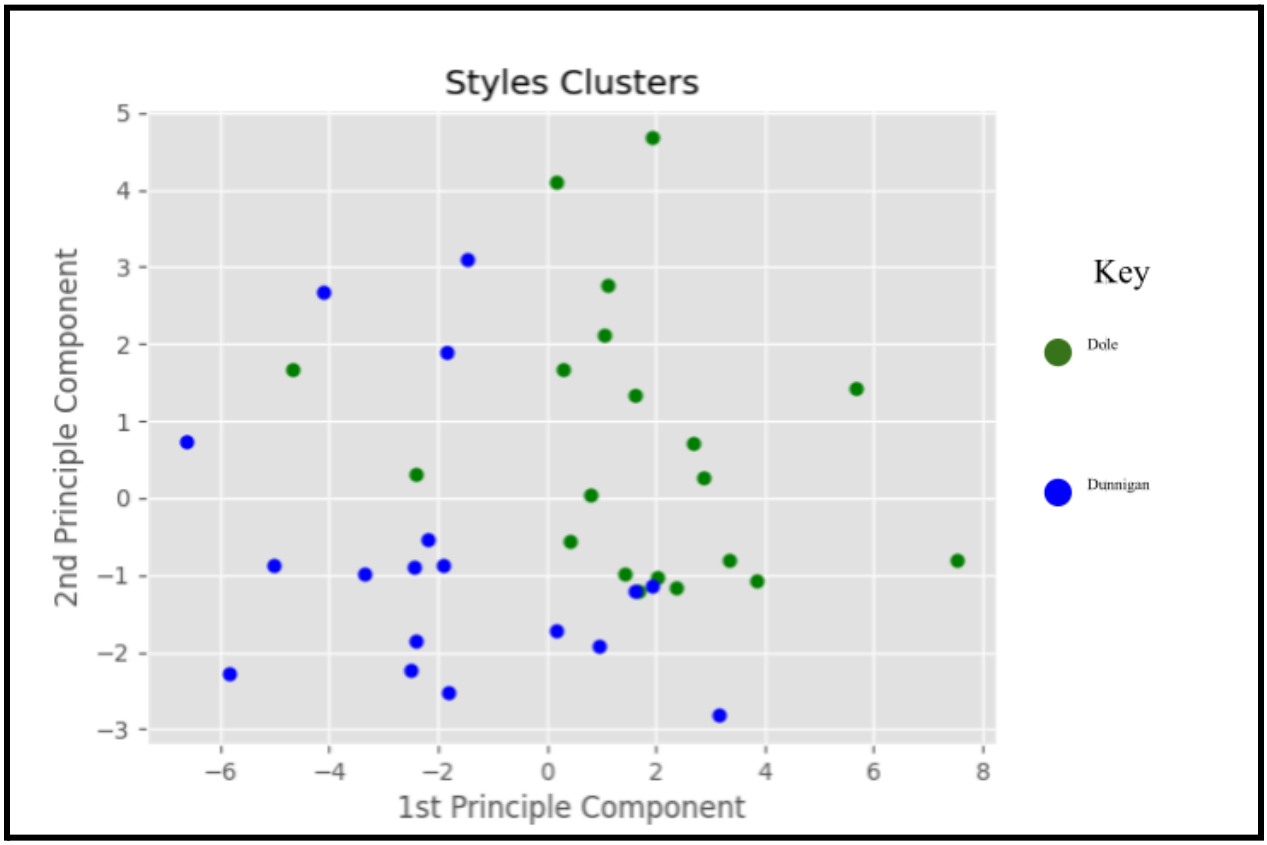
Appendix F.

PCA Graph of the Briggs and Volokhonsky Translations



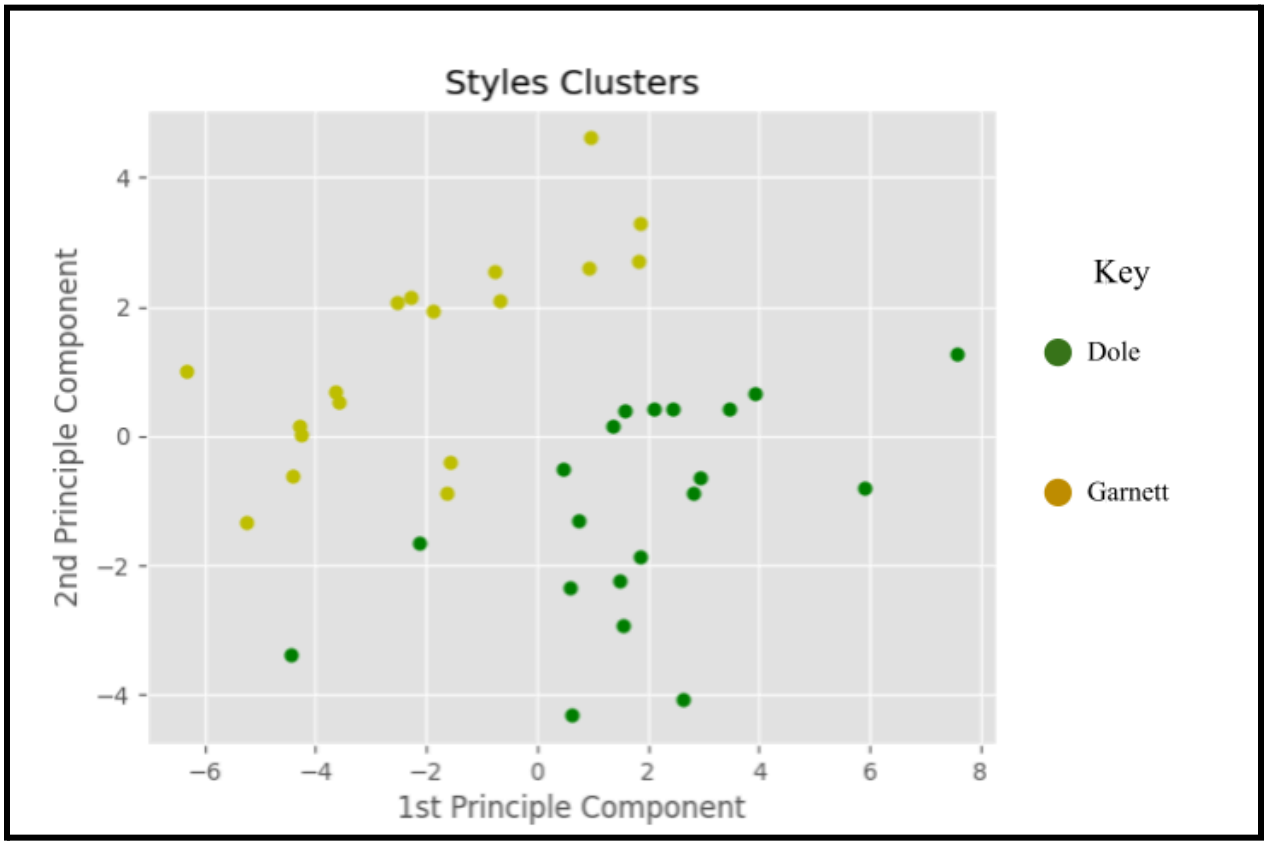
Appendix G.

PCA Graph of the Dole and Dunnigan Translations



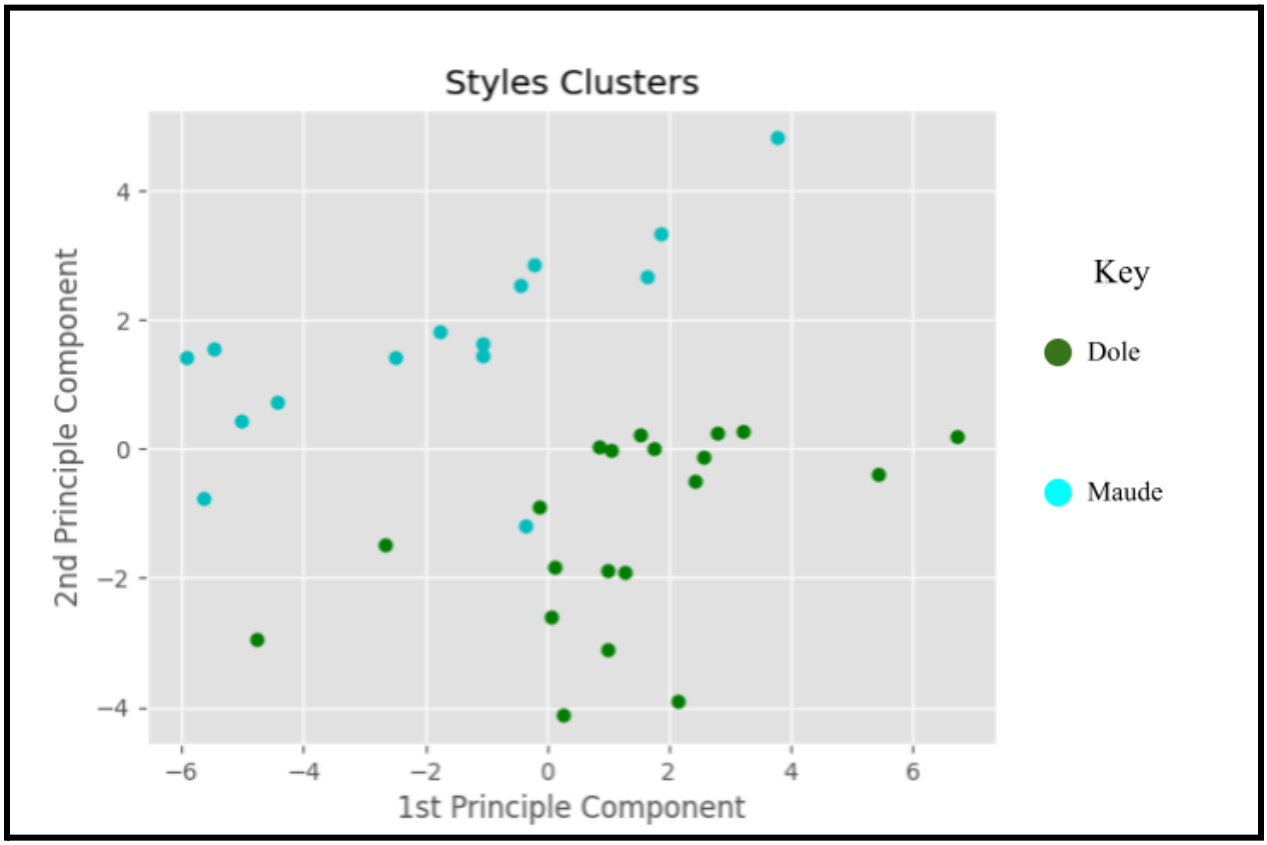
Appendix H.

PCA Graph of the Dole and Garnett Translations



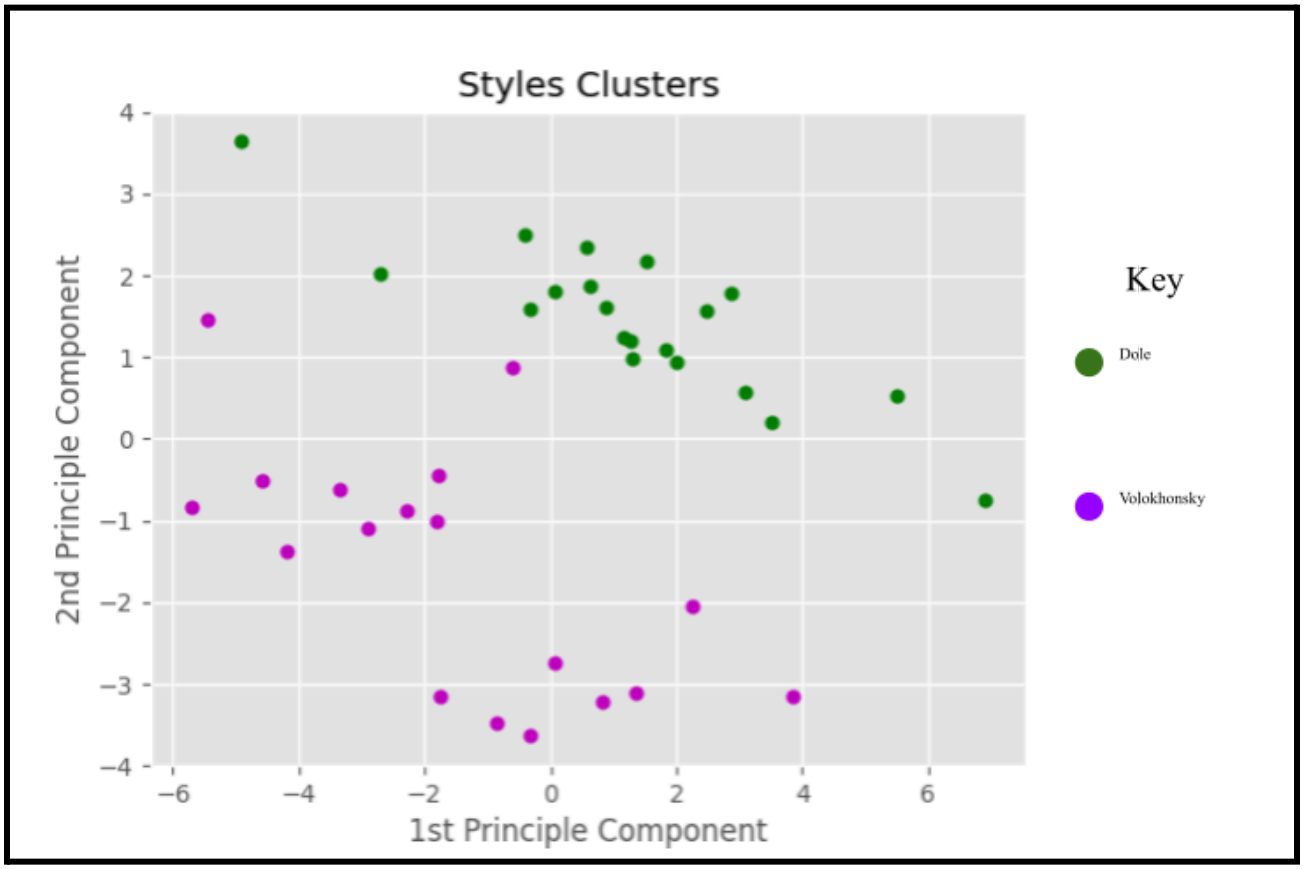
Appendix I.

PCA Graph of the Dole and Maude Translations



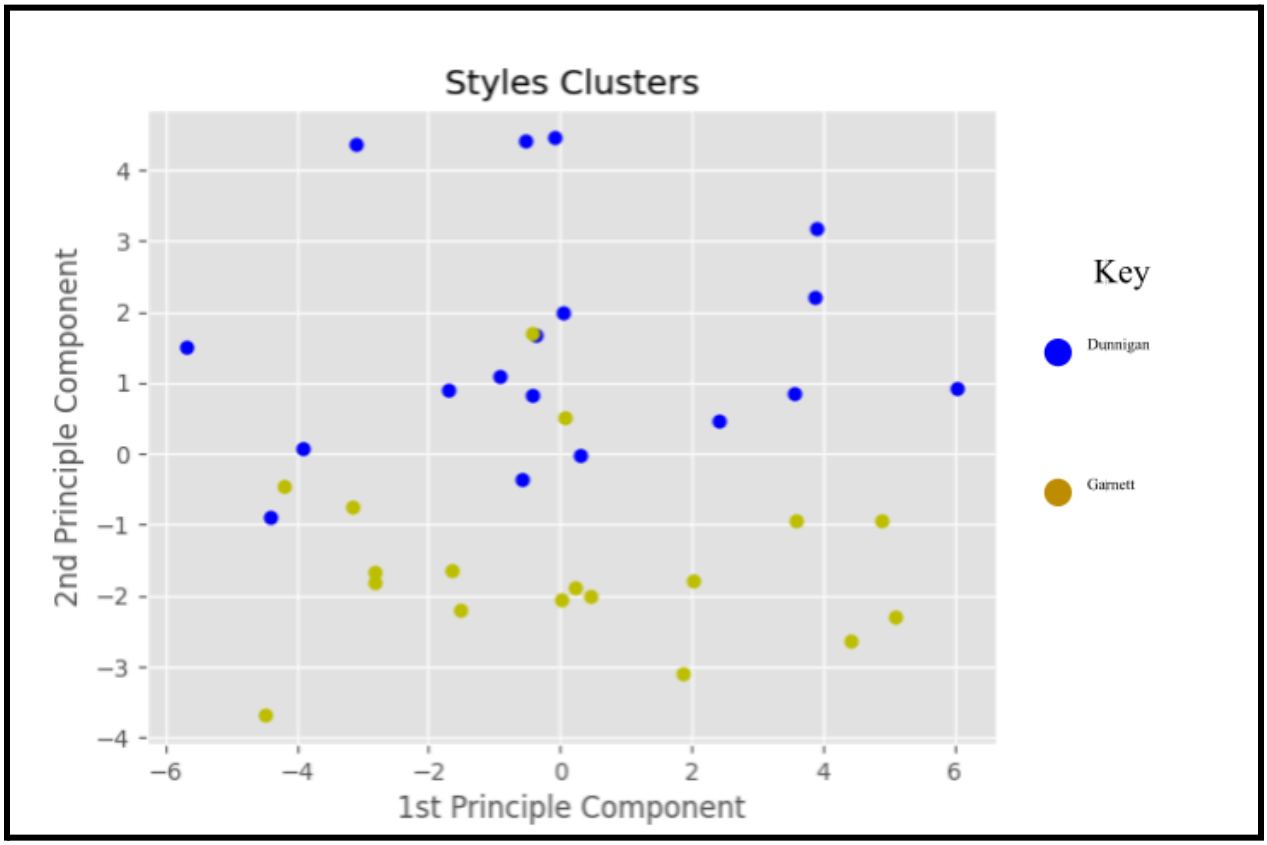
Appendix J.

PCA Graph of the Dole and Volokhonsky Translations



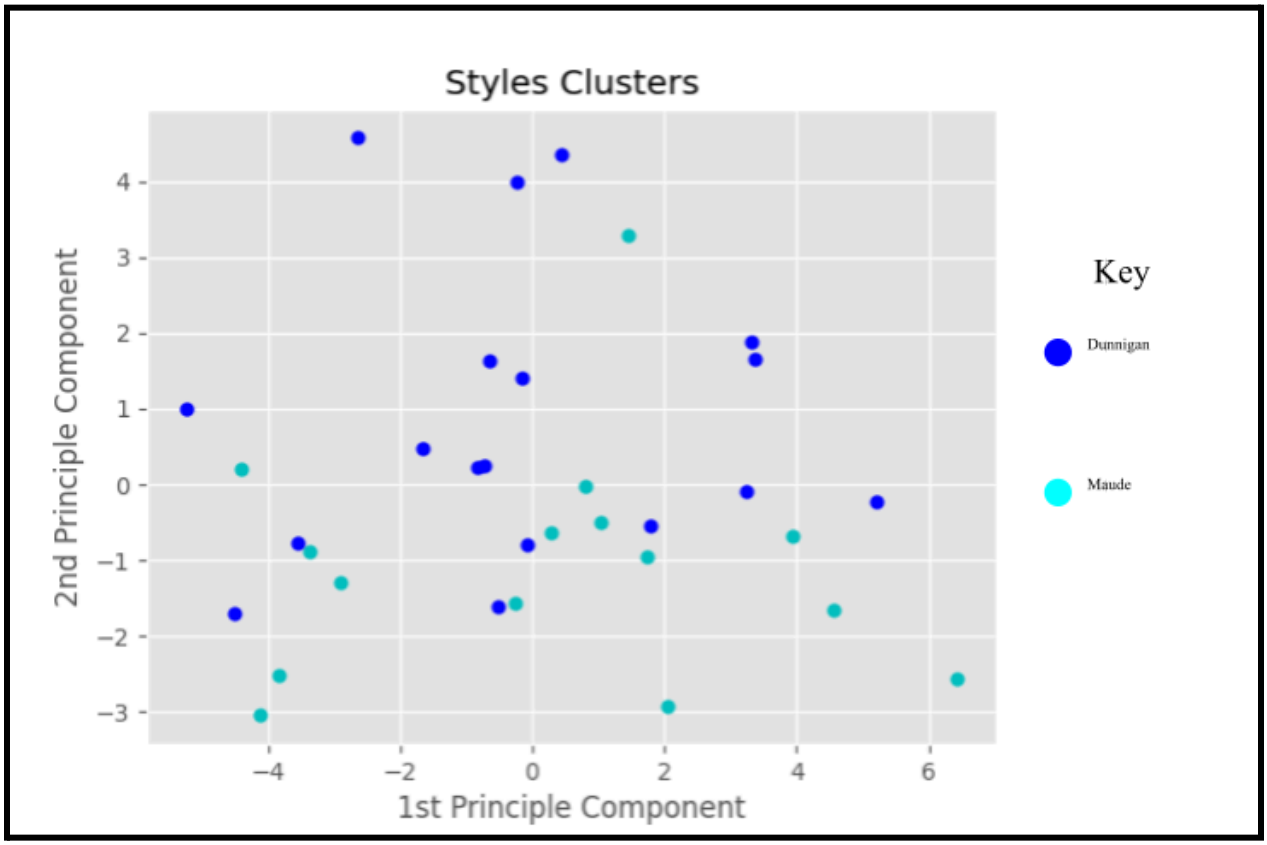
Appendix K.

PCA Graph of the Dunnigan and Garnett Translations



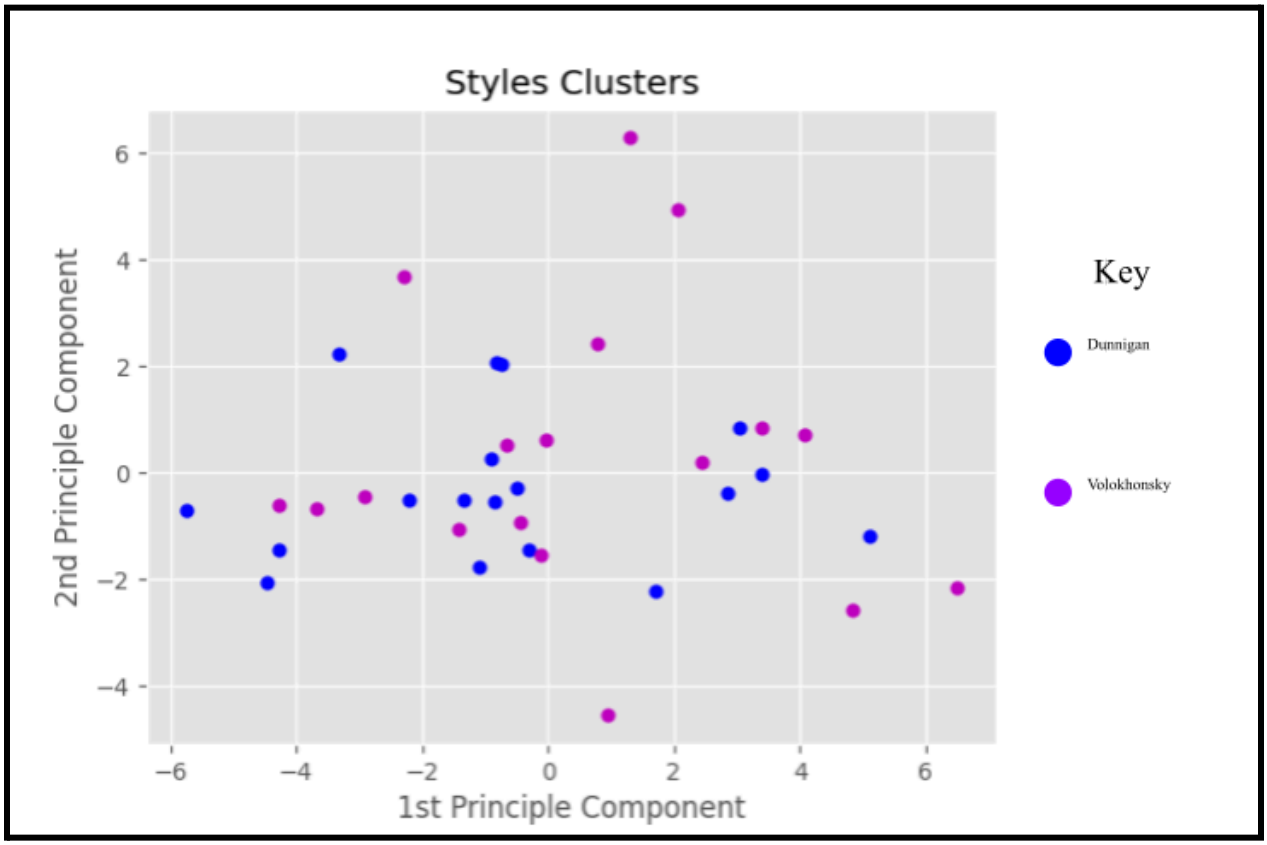
Appendix L.

PCA Graph of the Dunnigan and Maude Translations



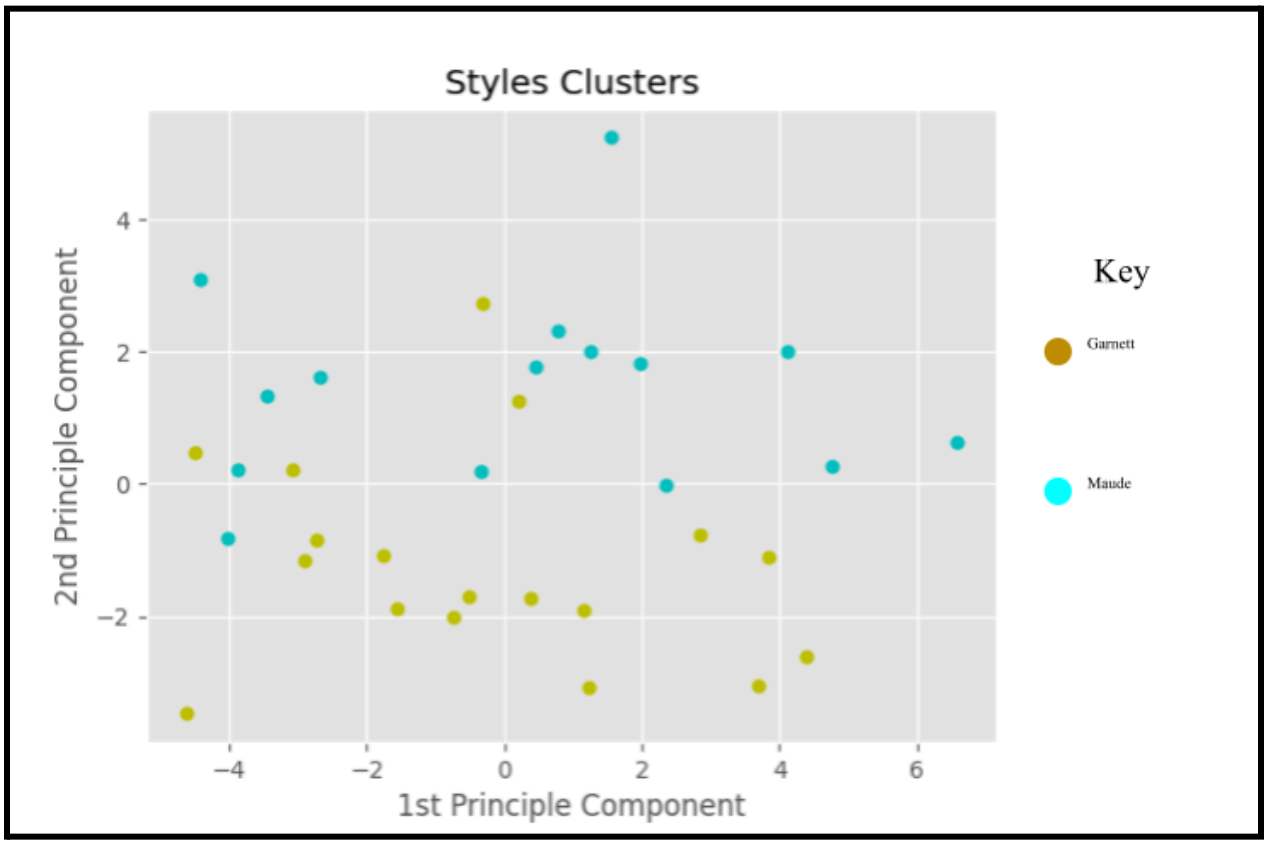
Appendix M.

PCA Graph of the Dunnigan and Volokhonsky Translations



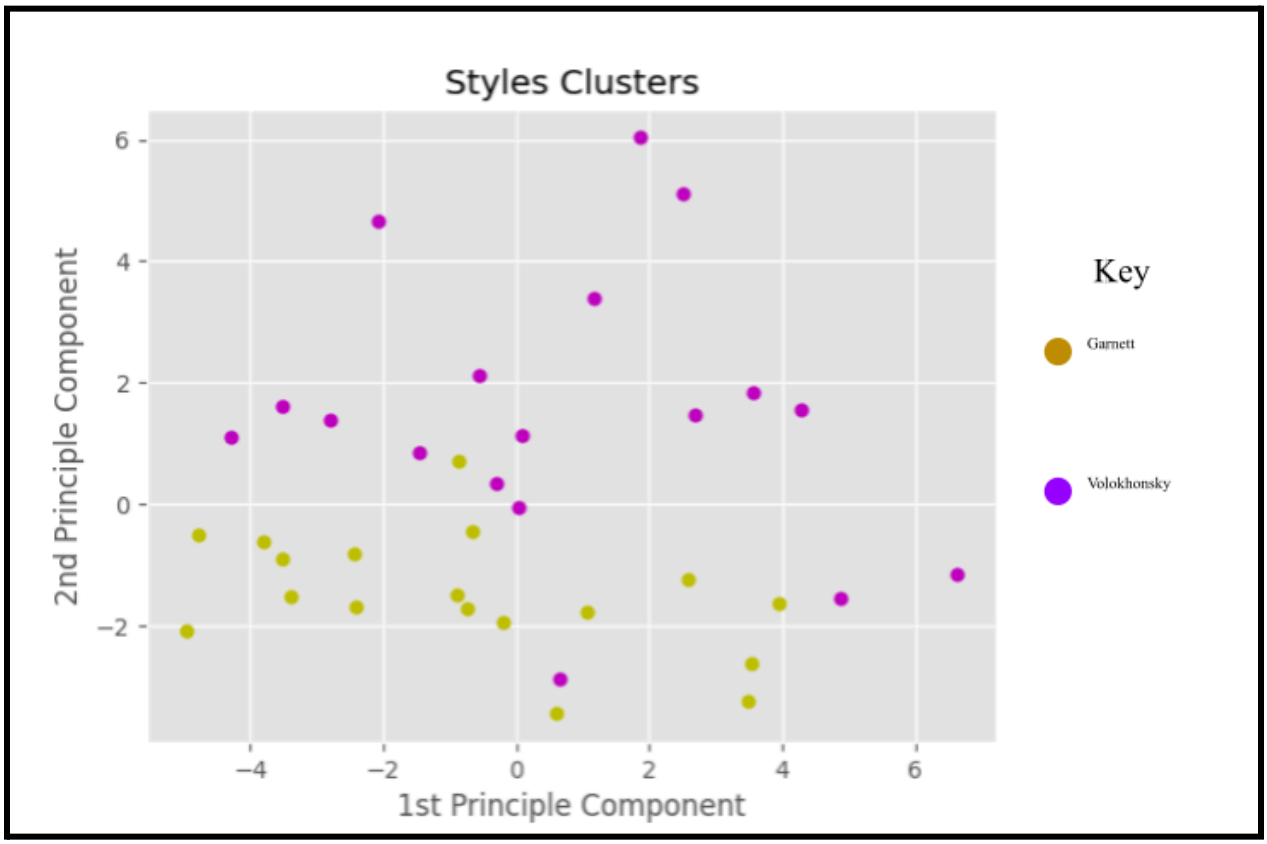
Appendix N.

PCA Graph of the Dunnigan and Volokhonsky Translations



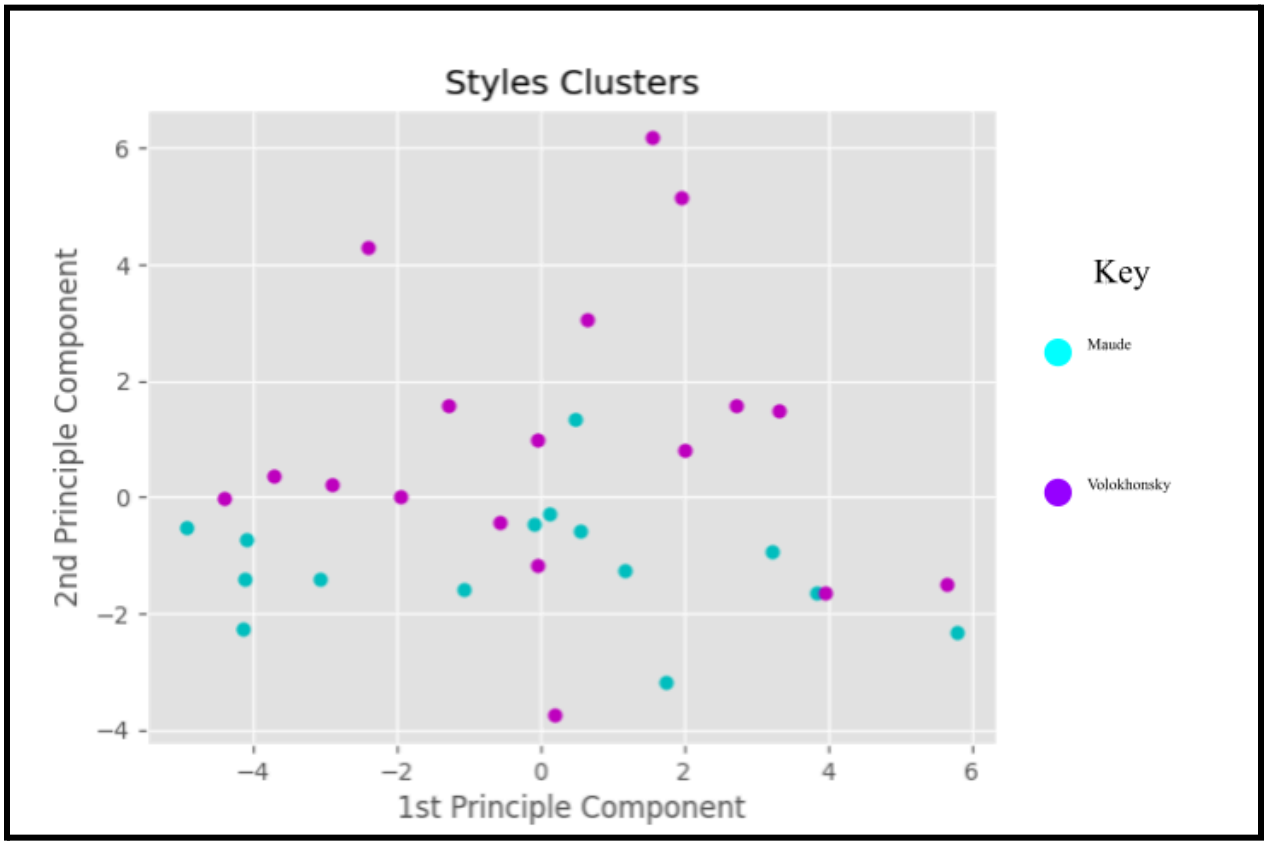
Appendix O.

PCA Graph of the Garnett and Volokhonsky Translations



Appendix P.

PCA Graph of the Maude and Volokhonsky Translations



Appendix Q.

A snippet of the algorithm written in the Python programming language

```
if __name__ == '__main__':

    # Create Styles Clusters Graph:

        #documents = ["text/example.txt"]
        #documents = ["text/knudsen.txt","text/shakespeare.txt"]
        documents = ["WAP/briggs.txt", "WAP/dole.txt", "WAP/dunnigan.txt", "WAP/garnett.txt", "WAP/maude.txt",
"WAP/volokhonsky.txt"]
        #documents = ["WAP/maude.txt", "WAP/volokhonsky.txt"]
        vector = []
        vectorList = []

        for document in documents:
                vector += FeatureExtration(open(document, encoding="utf8").read(), winSize=50, step=50) #change to 50/10
(winsize, step)
                vectorList.append(FeatureExtration(open(document, encoding="utf8").read(), winSize=50, step=50)) #change to 50/10
(winsize, step)

        colorArrangement = assignColors(vectorList)

        #ElbowMethod(np.array(vector))
        Analysis(vector, colorArrangement)
```

Works Cited

Agarwal, Shilipi. “Which War and Peace translation you should pick?” ThinkSync, https://thethinksync.com/2021/05/which-war-peace-translation-you-should-pick/. Accessed 16 Dec. 2023.

Battles, Paul. “Intertextuality and Sociolectal Differentiation in Old Saxon and Old English Verse: A Stylometric Analysis Using N-Grams.” Modern Philology, vol. 119, no. 4, May 2022, pp. 443–67. EBSCOhost, https://doi.org/10.1086/719239.

“Best English Translation of War and Peace?” Reddit, https://www.reddit.com/r/tolstoy/comments/11e23rm/best_english_translation_of_war_and_peace/. Accessed 19 Dec. 2023.

Biber D, S Conrad, G Leech, *The Longman student grammar of spoken and written English*, (Harlow: Longman, 2002). ISBN: 0 582 237262.

Brunet v. Le Vocabulaire De Jean Giraudoux : Structure Et évolution : Statistique Et Informatique Appliquées à L’étude Des Textes à Partir Des Données Du Trésor De La Langue Française. Le Vocabulaire des grands écrivains français (Genève, Slatkine, 1978). ASIN: B0000E99PZ.

Elahiand, Hassaan Haris Muneer. Identifying Different Writing Styles in a Document Intrinsically Using Stylometric Analysis. Zenodo, 3 July 2018, doi:10.5281/zenodo.2538334.

El-Fiqi, Heba, et al. “Network Motifs for Translator Stylometry Identification.” PLoS ONE, vol. 14, no. 2, Feb. 2019, pp. 1–33. EBSCOhost, https://doi.org/10.1371/journal.pone.0211809.

Esteves, Lenita. “Intellectual Property and Copyright: The Case of Translators.” Translation Journal, vol. 9, no.3, July 2005. Accessed 16 April 2024.

Fuggle, Lucy. “What’s the best translation of War and Peace by Leo Tolstoy?” Tolstoy Therapy, https://tolstoytherapy.com/best-translation-war-and-peace/. Accessed 16 Dec. 2023.

Hedegaard S, Simonsen JG. Lost in translation: authorship attribution using frame semantics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. vol. 2 of HLT’11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 65–70.

Honore A, Some simple measures of richness of vocabulary. Assoc. Literary Linguistic Comput. Bull. 7, 1979.

Juola, Patrick, "Authorship Studies and the Dark Side of Social Media Analytics," Journal of Universal Computer Science, vol. 26, no. 1, 2020, pp. 156–70. 10.3897/jucs.2020.009

Kestemont M. What Can Stylometry Learn From Its Application to Middle Dutch Literature? Journal of Dutch Literature. 2012;2(2):46–65.

Kumiko Tanaka-Ishii, Shunsuke Aihara; Computational Constancy Measures of Texts—Yule's *K* and Rényi's Entropy. *Computational Linguistics* 2015; 41 (3): 481–502. doi: https://doi.org/10.1162/COLI_a_00228

Mikhailov M, Villikka M. Is there such a thing as a translator's style? In: Rayson P, Wilson A, McEnery T, Hardie A, Khoja S, editors. Proceedings of the Corpus Linguistics 2001 conference. Lancaster: Lancaster University (UK); 2001. p. 378–386.

Mosteller and Wallace. Inference and Disputed Authorship : The Federalist. Addison-Wesley, Reading, MA. 1964.

Murphy, Mary J. "130 Years Ago: 'War and Peace' Finally Published in English." New York Times, https://www.nytimes.com/2016/01/29/arts/television/130-years-ago-war-and-peace-finally-published-in-english.html#:~:text=The%20novel%20was%20published%20in,%E2%80%9CA%20Famous%20Russian%20Novel.%E2%80%9D. Accessed 16 Dec. 2023.

Ríos-Toledo, Germán, et al. “Detection of Changes in Literary Writing Style Using N-Grams as Style Markers and Supervised Machine Learning.” PLoS ONE, vol. 17, no. 7, July 2022, pp. 1–24. EBSCOhost, https://doi.org/10.1371/journal.pone.0267590.

Rybicki J. The great mystery of the (almost) invisible translator: Stylometry in translation. In: Oakes MP, Ji M, editors. Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research. Studies in Corpus Linguistics. John Benjamins Publishing; 2012. p. 231–248.

Sichel, H. S. “On a Distribution Law for Word Frequencies.” *Journal of the American Statistical Association*, vol. 70, no. 351, 1975, pp. 542–47. *JSTOR*, https://doi.org/10.2307/2285930. Accessed 29 Apr. 2024.

Yule, G. U. The statistical study of literary vocabulary. Cambridge University Press, 2014.