Using Latin Treebank Data to

Study Authorial Stylistics

By Brian Clark

**Introduction to Project**

For this project, I have worked with digital tools to make book 13 of Ovid’s *Metamorphoses* as accessible as possible not only to students of Latin, but also to a wider less specialized audience. I have created a treebank, which will be explained in-depth later, of this book to provide a digital commentary on the Latin text. My translation alignment of the opening of this book compares the Latin text to an English and Spanish translation. By combining these tools, a reader will be better prepared to understand this portion of the *Metamorphoses*, and engage directly with the original text on a deep level.

Additionally, a goal of this project was to work with the data generated by the treebank edition of Ovid[[1]](#footnote-1). One aspect has been to compare authorial stylistics between different authors, which will be elaborated later. I have also written a program that allows teachers to make use more efficient use of treebanks when looking for example sentences.

**Book 13**

Book 13 was chosen for two main reason – its subject matter and its length. Book 13 begins with two length speeches from Ajax and Ulysses in their contest for the arms of Achilles. Then, Ovid relates a series of transformations connected to the aftermath of the fall of Troy, such as Ajax, Polyxena, Hecuba, and Memnon. He goes on to offer the start of his “Little *Aeneid*” and closes the book with Polyphemus’ pursuit of Galatea and its consequences. The connection to the *Iliad* and the wider Epic Cycle is significant because book 13 covers myths that most audiences would have some level of familiarity with, but they are rarely exposed to a continuous narrative of these events, such as the one Ovid offers. Choosing a book that is connected to more commonly known Greek mythology provides the chance that this project will reach a wider audience than if the project had been based on more obscure heroes or stories.

This book was also specifically chosen with high school students in mind. A common advanced Latin class in a high school curriculum is the Advanced Placement Vergil course. This course focuses on selections from the first six books of the *Aeneid* in Latin and the entirety of the work in English. After students have taken the AP Latin course on Vergil, working with book 13 would be an engaging way to revisit the journey of Aeneas and other epic themes. While Ovid’s so-called “Little *Aeneid*” is told in various sections of the last three books of the *Metamorphoses*, not only does it begin in book 13, but the portions of the *Aeneid* that are retold here are found on the Latin portion of the AP syllabus. Students would already be familiar with Aeneas and his journey to Delos and other islands before landing at Troy. With that background, they would be more prepared to study a different Augustan era poet who uses the same meter and covers the same subject matter. A comparison of the characterization of Aeneas in Ovid and Vergil would be way to allow the students to revisit Latin they have already read and reread it in the context of Ovid’s intentionally reworked version.

Another reason why this book was selected was because of its length. At 968 lines, book 13 is the longest within the *Metamorphoses*. In fact, is the longest by about one hundred lines. The aim was to choose a book that was large enough to work with when analyzing the XML data produced by the treebank. With these extra hundred lines, the treebank is roughly fifty sentences longer, which is not an insignificant increase. As analysis on treebank data is relatively new and understudied, there is not yet data on the number of sentences that are needed to produce meaningful data. Without an established threshold, I have erred on the side of having more data, whenever possible.

**Treebanks**

A treebank is a digital commentary that represents every semantic and syntactic connection in a text. For each sentence in a corpus, the user identifies the morphology of every word in order to disambiguate the correct form from a form with a different case or tense. The user also links each word to what the word that it modifies and labels its grammatical function in the sentence. The user must account for every word, clause, and even punctuation mark and understand its function in the sentence.

Working with treebanks has many pedagogical benefits. First, students can be assigned to read a portion of a text that has been treebanked and can consult the treebank when they do not understand how a word or phrase is functioning in the sentence. With every word fully identified in the treebank, students can quickly and accurately resolve their doubt as to how any part of the sentence functions as well as what units should be taken together. Learning to read from a treebank will increase the speed with which upper level students read, once they have built a solid knowledge of the language’s grammar. As a result, students can be assigned more material to read or to review for an exam if these passages have been treebanked.

The process of creating their own treebank is also invaluable to students. A treebanking assignment ensures that the student accounts for every word in the sentence and to understand every connected idea. This tests the student’s knowledge of morphology, grammar, and their understanding of the sentence. This close-reading approach is useful for students still attempting to master complex morphology or grammar, as well as students who are more advanced. Being held accountable for what units are being coordinated and specific functions of various clauses helps the student develop a deep understanding of how a sentence works. Repeated mistakes that arise in a group of students’ assignments, such as confusing indirect speech with a nominal substantive clause, will guide the teacher in areas that need more review.

Assigning students a portion of text that has already been treebanked allows them or the teacher to compare the students’ work to a published, corrected treebank. This practice of comparison and correction trains the student to produce more accurate treebanks and gives them confidence in reading a treebank. Once they have become familiar with and confident working with treebanks, they can then be assigned a portion of text that has not been treebanked before. Once these assignments are haven completed and corrected, they can be published for future students to use.

**Alignment**

In addition to the treebank of book 13 of the *Metamorphoses*, I have also completed a tri-lingual translation alignment of Ajax’s speech, which opens the book. This speech spans the first 122 lines and proved to be a manageable selection to align. My alignments include the Latin text from the Latin Library, the English translation done by A. S. Kline for *Poetry in Translation[[2]](#footnote-2)*, and Ana Perez Vega’s Spanish translation for the Biblioteca Virtual Miguel de Cervantes[[3]](#footnote-3). This limited passage was also chosen to comply with copyright laws, which allow for quotation of a set amount of non-public domain work for critique and evaluation. In its essence, alignment is evaluation of the level of closeness found in a translation when compared to the original text.

In order to approach my alignment, I consulted with several scholars working on translation alignment, most notably Maryam Foradi, Tariq Yousef, and Chiara Palladino. Through my consultation with them, I have developed editorial standards for translation alignment. As the practice of translation alignment has been emerging and becoming more prominent, set editorial standards have generally been lacking. The aim of set standards is to regularize the level of individual interpretation that is often overlooked in alignment. While my standards are written with a Latin-English-Spanish alignment in mind, it is my hope that they serve as a model and a base for future scholars. While certain aspects are specific to my languages, such as aligning the personal “a” in Spanish, many of my standards can be applied to any language.

The aim of my alignment has been to align as much as possible, in order to allow the reader to comprehend the full extent of the translation. A strict one-to-one translation, while helpful for automated vocabulary tests and comparisons, does now allow the reader to explore the degree of liberty granted to each translator. My goal has been to give my reader a view into the different ways that Latin can be rendered in modern languages. Consequently, I have intentionally aligned passive verbs with active verbs, singulars with plurals, and other items that are strictly incongruent, but express a similar meaning. The practice of the “royal we” is a useful example for the importance of this ideal. While Latin readily uses plural forms when referring to a singular speaker during a rhetorical speech, modern languages do not maintain this practice to the same degree. This practice is used at times, but with a different connotation. Still, there is enough overlap in the meaning that one should align an English “I” or Spanish “yo” to a Latin “nos” with the understanding that the sense is the same.

My editorial standards also argue strongly for aligning certain sense units as multi-word phrases, rather than on a word-by-word basis. For example, the Latin preposition “in” has a variety of meanings ranging from “in”, “into”, “against”, and many more. To align the Latin “in” directly to these English words is accurate, but does it strips the word of its sense when combined with the word that follows it. To align the Latin phrase “in classem” to the English phrase “against the fleet” not only preserves the semantic connection between these words, but allows a reader to more efficiently make use of the translation alignment. There is no one translation of many Latin prepositions, so aligning them as phrases provides a more complete translation.

The purpose of this alignment was to compare recent translations of the *Metamorphoses* for their accuracy to the original text, as well as to demonstrate the way in which translation and grammatical constructions varies across language. The modern languages that I used were chosen, beyond being languages that I am fluent in, for their wide use in the United States. In addition to the treebanks, future students can make use of this translation alignment in order to better understand this passage of Ovid’s *Metamorphoses*. With a growing population of Spanish-speakers in the United States, as well as Spanish being the predominate foreign langue taught in schools, an approach the unites Latin and Spanish will be increasingly important and enlightening for the students. An alignment that compares these two languages is an effective tool to draw concrete parallels between grammatical structures and vocabulary items that are shared by both languages. A certain percentage of students study Spanish in elementary and middle school, and then go on to study Latin later in life. With this grammatical knowledge of Spanish, a tri-lingual alignment of Latin-English-Spanish will allow the students to utilize their familiarity with these languages to better access the Latin.

**Analysis Introduction**

For the analysis portion of this project, the research question was to use the treebank data to analyze the authorial stylistics of Ovid’s *Metamorphoses* and Vergil’s *Aeneid*. These authors were selected because they both wrote epic poetry during the Augustan Age. Comparing two authors in the same genre is a logical first comparison because it highlights the differences between the authors, rather than differences that arise due to the change in genre. A study of a prose work and a poetic work would likely show a great disparity, due to the different constructions that are popular in each. Even poetic works of different genres are expected to be different when analyzed together, due to the nature of a meter like dactylic hexameter and elegiac couplets, to give an example. One future path that would be interesting would be to compare works of different genres written by the same author, in order to see to what degree the author varies his style across different types of writing. For example, a study of Ovid’s *Metamorphoses* and his *Amores* or *Heroides* would likely be revealing.

Another reason that these authors and texts were chosen was because of the material that had been treebanked previously. Once I completed my treebank of book 13 of the *Metamorphoses*, I concatenated that data with the treebank for book 1 on the *Metamorphoses.* This produced a file with 790 sentences over 1747 lines. For Vergil, I worked with the treebanked AP syllabus, as well the entiery of book 1. Combined, the analysis was done on 586 sentences of Vergil. In order to manage the size difference in my data sets, all of my analysis was run on percentages based on number of words in each file. Basing my percentages on number of words was more accurate than using the number of lines, given the varying lengths of a hexameter and the handful of lines left unfinished by Vergil.

In order to put these results in context, I analyzed Satire 6 by Juvenal and book 1 of Caesar’s *De Bello Gallico*. Having more data allowed me to distinguish results were significant and results that were due to similar meter or other factors. Additionally, I treated books 1 and 13 of Ovid separately, in order to see to what degree his style varies between these books. While it is possible that Ovid did vary his style based on thematic content, it is assumed that no one book is drastically different from another of the same work. If a test were to reveal stark differences between book 1 and book 13 of the *Metamorphoses*, then there would likely be a problem with the analysis.

**Teaching Tool**

In addition to the analysis, part of this project was to use the treebank XML data to make meaningful tools for teachers to better make use of treebanks in their classroom. As mentioned above, the pedagogical benefits of having students create their own treebanks and the speed that using treebanks as reading aids allows are valuable both in a classroom and for independent study. I have written a program using of the XML data that allows the user to search for specific grammatical features in their treebank data file. This program is useful for finding passages within a given author that utilize the grammatical concept that is being emphasized in the current lesson. If one searches for accusative direct objects or nominative subjects, the program will logically return many sentences, since these constructions appear frequently. However, there are constructions that are used much more infrequently, to the point that it can often be hard to find examples of these. When the user searches for ablatives of degree of difference or genitives of material, the program will return far fewer sentences, which reflects the scarcity of these constructions in most authors.

The benefit of this program is that is allows teachers to pull authentic examples from Latin literature for any given grammatical construction. They can either augment textbook sentences with these passages from Roman authors, or, if they are aiming to omit sentences written by modern scholars all together, they can give their students these sentences without alterations. In the push to make sure that students are exposed to as much of the language in its original form as possible, this program not only makes it easier for a teacher to find examples, but also potentially removes the need for sentences written for modern textbooks.

This program is an improvement upon searching through the treebank XML data by hand, and should be seen as more than just a look-up function. By printing not only the complete sentence, but also the passage reference for the sentence, the user is more able to make efficient use of the output of the program. They do not have to search the XML by hand and reconstruct the sentence from the “form” attribute on every “word” element. This step is complicated by the enclitics such as “–que” or “–cum” that become distinct words in the data file, as well as the words that have been added in parentheses to complete ellipses. My program recombines the enclitics and ignores added elliptical words, so that it can print the sentence in its original form.

Additionally, the program frees the user from the need to switch between the data file and another edition of the text using the passage reference alone. By returning both the passage reference and the sentence without any editorial additions or divisions, the user can make immediate use of the output that they receive for any given grammatical construction. The sentences can be copied directly into worksheets or assignments for students, or the line reference can be put directly into the Perseus Scaife Viewer[[4]](#footnote-4) to focus on that specific passage.

**Analysis & RESULTS**

As mentioned above, the following analysis was done on books 1 and 13 of Ovid’s *Metamorphoses*, selections of Vergil’s *Aeneid* that are comparable in length, Satire 6 of Juvenal, and book 1 of Caesar’s *De Bello* Gallico. My first program extracted all of the relation tags used in a specific file and returned the percentage of that tags use. This was done by dividing the number of occurrences of that tag divided by the number of words in the file excluding punctuation characters and elliptical words[[5]](#footnote-5). Then, I analyzed the correlation and R-squared in order to find the meaningful connection between the various authors’ stylistics. Figure 1 shows the statistical data for this analysis, and Figure 2 shows a graph of the percentages.

**Figure 1: All Tags**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.953322208 | 0.986362133 | 0.984920375 | 0.8967459 |
| Ovid 13 |  |  | 0.951722435 | 0.937126072 | 0.888391787 |
| Vergil 1 |  |  |  | 0.975131233 | 0.901355661 |
| Juvenal 6 |  |  |  |  | 0.884098007 |
| Caesar 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.908823233 | 0.972910258 | 0.970068145 | 0.80415321 |
| Ovid 13 |  |  | 0.905775594 | 0.878205275 | 0.789239967 |
| Vergil 1 |  |  |  | 0.950880922 | 0.812442028 |
| Juvenal 6 |  |  |  |  | 0.781629285 |
| Caesar 1 |  |  |  |  |  |

The results of this analysis did not reveal any significant difference between these authors. Caesar is less connected to the other texts, with an R-squared around 0.80 when compared to the other texts. The remainder of the texts all seem to be highly connected, with R-squared figures ranging from 0.97 to 0.87. As one would expect, given the fact that Caesar is the only prose author in this study, his data stands out, but not to a significant enough degree to be conclusive. This data supports the idea that a more targeted analysis is needed to pinpoint meaningful difference in authorial stylistics. Analyzing the percentage of tags used does not provide a narrow enough lens through which the individual author’s style becomes clear. This analysis does suggest that there is a specific way that elite literary Latin is written. Generally speaking, when looking at every tag used, there is not enough deviation to suggest that the connection is due to anything more than the fact that these are all texts written in Latin.

In order to look more closely at the differences between authors, the next round of analysis focused on a specific tag. I began looking at ablative absolutes, and what each author uses to modify them. Ablative absolutes make up less than two percent of a given work, based off of the previous analysis, and therefore proved to be a meaningful tag to start with. As shown in the statistics in Figure 3, Caesar’s use of modification with ablative absolutes is radically different from the other authors. With his highest R-squared being 0.11, this test clearly demonstrates the different style used between these authors. Caesar seldom leaves an ablative absolute unmodified, with only 2% of them without some other word in book 1, compared to the other authors who use unmodified ablative absolutes at least a third of the time, or Vergil who uses them 83% of the time.

**Figure 3: Modification of AB-ABSOL**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.886275608 | 0.798983971 | 0.859593595 | 0.169100281 |
| Ovid 13 |  |  | 0.944942597 | 0.93138137 | 0.344953977 |
| Vergil 1 |  |  |  | 0.944817552 | 0.285595217 |
| Juvenal 6 |  |  |  |  | 0.095483571 |
| Caesar 1 |  |  |  |  |  |
| Caesar 8 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.785484453 | 0.638375386 | 0.738901149 | 0.028594905 |
| Ovid 13 |  |  | 0.892916511 | 0.867471257 | 0.118993246 |
| Vergil 1 |  |  |  | 0.892680207 | 0.081564628 |
| Juvenal 6 |  |  |  |  | 0.009117112 |
| Caesar 1 |  |  |  |  |  |
| Caesar 8 |  |  |  |  |  |

While Caesar is the clear outlier, the other authors are relatively highly connected, without any pair standing out to a significant degree. However, book 13 of Ovid and the data for Vergil do have an R-squared of 0.89. This is not overly compelling taken alone, but it is noteworthy that book 13 is more connected to Vergil than book 1 is. Book 13 of the *Metamorphoses* contains the fall of Troy and the wanderings of Aeneas, as does the *Aeneid*. This connection, when bolstered by other analysis, may show a link between these passages.

Analysis done on modification of direct objects was far less conclusive than the same analysis done with ablative absolutes. Like the analysis done with every tag used by an author, modification of direct objects reveals that this construction functions similarly throughout Latin. Direct objects make up 7% of texts on average, which appears to be too high to be meaningfully significant. Figures 5 and 6 show the data for modification of direct objects. Even the high connection between Ovid 13 and Vergil does not prove to be significant, given the other high connections seen with various authors.

**Figure 5: Modification of A-DO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.992636193 | 0.986377393 | 0.991739424 | 0.990207146 |
| Ovid 13 |  |  | 0.99830929 | 0.971935152 | 0.97311223 |
| Vergil 1 |  |  |  | 0.960169235 | 0.961170082 |
| Juvenal 6 |  |  |  |  | 0.994541988 |
| Caesar 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.985326611 | 0.972940361 | 0.983547086 | 0.980510192 |
| Ovid 13 |  |  | 0.996621438 | 0.94465794 | 0.946947412 |
| Vergil 1 |  |  |  | 0.92192496 | 0.923847927 |
| Juvenal 6 |  |  |  |  | 0.989113766 |
| Caesar 1 |  |  |  |  |  |

Next, I did three different analyses on the words that mark coordination, prepositions, and subordinate clauses. For coordination, asyndeton was calculated for any “word” element with a “COORD” value for the “relation” attribute that had punctuation as its “lemma” attribute. The differences largely were found in the distinction between prose and poetry. Figure 7 shows that the Caesar material is less significantly connected than most of the other texts, although there are some interesting figures with Juvenal 6.

**Figure 7: Words used to show coordination**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.987430136 | 0.989897157 | 0.884470364 | 0.833343687 |
| Ovid 13 |  |  | 0.970102974 | 0.831122848 | 0.767321471 |
| Vergil 1 |  |  |  | 0.922879158 | 0.891038965 |
| Juvenal 6 |  |  |  |  | 0.983778654 |
| Caesar 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.975018273 | 0.979896381 | 0.782287825 | 0.694461701 |
| Ovid 13 |  |  | 0.94109978 | 0.690765188 | 0.58878224 |
| Vergil 1 |  |  |  | 0.85170594 | 0.793950438 |
| Juvenal 6 |  |  |  |  | 0.967820441 |
| Caesar 1 |  |  |  |  |  |

Figure 8 shows each author’s inclination to use certain coordinators over others. Caesar favors using *atque* and he and Juvenal both favor *et*, while Ovid and Vergil make much more frequent use of *que*. The use of asyndeton appears to be relatively consistent across these authors, at around 18.5%.

The results of this analysis done on preposition words was surprising. Book 1 of the *Metamorphoses* stands out as significantly different from the other works. As seen in Figure 9, the data shows this book as notably separate from the rest. While the R-squared for the other works ranges from 0.88 to 0.72, the comparisons to Ovid 1 range from 0.42 to 0.19. It is unclear what about this book of Ovid causes it to depart to such a degree from the other works.

Figure 9: Words for prepositions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.527901302 | 0.65505057 | 0.438012041 | 0.54926937 |
| Ovid 13 |  |  | 0.859127662 | 0.943138648 | 0.892809857 |
| Vergil 1 |  |  |  | 0.848702879 | 0.856822262 |
| Juvenal 6 |  |  |  |  | 0.92258545 |
| Caesar 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.278679785 | 0.42909125 | 0.191854548 | 0.301696841 |
| Ovid 13 |  |  | 0.73810034 | 0.889510509 | 0.79710944 |
| Vergil 1 |  |  |  | 0.720296577 | 0.734144388 |
| Juvenal 6 |  |  |  |  | 0.851163913 |
| Caesar 1 |  |  |  |  |  |

Lastly, I analyzed the words that trigger subordinate clauses, labeled as “AuxC” in the “relation” attribute in the treebank XML. In this analysis, Juvenal stood out as significantly different from the other authors. Caesar also departed from Ovid and Vergil, but not to the same degree that Juvenal did. In figure 11, the division between the Augustan poets and the other authors is clear.

**Figure 11: Words for Subordinate Clauses**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.91518196 | 0.896372956 | 0.544620379 | 0.812635761 |
| Ovid 13 |  |  | 0.917617383 | 0.702648651 | 0.789362594 |
| Vergil 1 |  |  |  | 0.726184079 | 0.713744319 |
| Juvenal 6 |  |  |  |  | 0.63890233 |
| Caesar 1 |  |  |  |  |  |
|  |  |  |  |  |  |
| R-Squared |  |  |  |  |  |
|  | Ovid 1 | Ovid 13 | Vergil 1 | Juvenal 6 | Caesar 1 |
| Ovid 1 |  | 0.83755802 | 0.803484477 | 0.296611357 | 0.660376881 |
| Ovid 13 |  |  | 0.842021661 | 0.493715127 | 0.623093305 |
| Vergil 1 |  |  |  | 0.527343317 | 0.509430953 |
| Juvenal 6 |  |  |  |  | 0.408196187 |
| Caesar 1 |  |  |  |  |  |

The average R-squared for Juvenal is 0.43 and is 0.55 for Caesar. This stands in contrast to the data for Vergil and Ovid, which are all above 0.80. Here, the data suggests a connection in style between the Augustan poets when compared to authors of a different period.

**Analysis Conclusions**

The analysis has proven that an approach that is more nuanced than studying of the tags used by an author. In order to show authorial stylistics meaningfully, one must look at specific usages of various grammatical constructions. A broad view is able to show the distinction between poetry and prose, but there is no great significant difference among these texts. The connection is likely because these texts are all elite literary works all produced in Latin. With that in mind, a study of the same genre in both Greek and Latin would be interesting to see.

An analysis of modification of ablative absolutes also shows the divide between poetry and prose, but also reveals a deeper connection between specific authors. The higher connection of Ovid 13 to Vergil than Ovid 1 here is interesting due to the thematic connection between these works. With more evidence like this, one could confidently make an argument for Ovid mimicking a Vergilian style when he begins his “Little *Aeneid.*” For now, it is an important connection, but merely one piece of data. Other analysis done on what modifies certain tags was less revealing. This analysis run on direct objects showed that this construction largely functions the same across these authors. Given the modest sample set, any future data that deviates from this trend would be significant.

When looking at what words are used for specific relations, the author’s unique choices become clear. One can see words that might be deemed as poetic or words that are more widely used in prose. Analysis done on words that mark coordination, prepositions, and subordinate clauses revealed more about individual authors or books being outliers, rather than wider significance.

**Limitations**

There are several limitations to this project. Chief among them is the sample size of the data. A logical critique is that this project requires more data from the *Aeneid* in order to make the size of the data sets closer to being equal. Ideally, this project would contain a great deal more data for both authors, not just Vergil. Running the analysis on entire works should not be necessary to yield meaningful results, but more data would increase the accuracy and persuasiveness of my findings. I am limited by the amount of each author that was already treebanked, which authors had substantial sections treebanked, and the time that it took to produce a treebank and review it for publication. In addition to completing the treebank for book 13 of the Metamorphoses, I also reviewed other previously treebanked passage from the *Aeneid* in order to add more data to that aspect of the project. Still, rerunning this analysis once more treebanks have been published will no doubt yield compelling and revealing results.

Another significant limitation of this project is my own coding abilities. While I was successful in gathering meaningful data from the treebank XML files, I was unable to do all that I had hoped. While I do not currently possess the requisite coding knowledge to analyze certain features, such as average lengths of subordinate or relative clauses, it is my hope that my work has laid the groundwork for future students. Additionally, as I plan to use treebanks in my future classroom, I will continue to work on improving and developing my code in order to access the full array of information that can be derived from the treebanks.

All worthwhile Digital Humanities projects are fundamentally collaborative endeavors that make use of a team with diverse skill sets. In an ideal world, I would have worked with a team that had collaborators with more advanced programming experience as well as those with knowledge of statistics. This environment would have contributed greatly to the final product of my project in terms of the querying and analysis of data.

**Future steps**

There are a variety of steps that both myself and others can take to continue this project. Perhaps the most straightforward and most inevitable is the production of more treebank data, particularly from the *Metamorphoses* and the *Aeneid*. Students have already completed a first pass of book 2 of the *Metamorphoses* and other sections of the *Aeneid*. Once they have been corrected and published, they can be incorporated into my analysis by concatenating these new files with my existing data sets, and rerunning the analysis. In the future, given enough data, my analysis can also be run on other authors and works beyond the two that I have analyzed here.

As mentioned above, someone with more sophisticated programming knowledge will be able to expand my work immensely. Because every word and clause is identified to such a high level of specificity, there is a wealth of knowledge contained in the treebank data. Average lengths of certain types of clauses are an area that would likely be revealing to analyze. Patterns in word order, such as when specific constructions tend to appear in conjunction to one another would also be an interesting avenue of exploration. I have not dealt with the morphological information provided by the treebanks at all in this project, and this is another important area for further work. As more data is published, scholars with a substantial coding background will be able to use my work as a jumping off point into exciting and fruitful research on countless philological topics.

Ultimately, the most fruitful analysis of treebank data to study authorial stylistics would use a wide array of the test that I have demonstrated here, and many others. These results would be marked as difference features, the combination of which would likely give far more specific insight into individual styles. This would be beneficial for different authors, but would likely be revealing for comparing works from different time periods or works from different genres. Being able to quantify how elegiac poetry differs from epic poetry will no doubt lead to greater philological insight into both genres.

1. This data is written in extensible markup language (XML) [↑](#footnote-ref-1)
2. <https://www.poetryintranslation.com/PITBR/Latin/Metamorph13.php> [↑](#footnote-ref-2)
3. <http://www.cervantesvirtual.com/obra-visor/metamorfosis--0/html/ff8ccec6-82b1-11df-acc7-002185ce6064_14.html#I_37_> [↑](#footnote-ref-3)
4. <https://scaife.perseus.org/> [↑](#footnote-ref-4)
5. Words need to be added into the text on occasion to complete the ellipsis, while still allowing the tree to hang properly. These words are also marked with an “artificial” attribute whose value is “elliptical”. Consequently, the program omits words with this attribute. [↑](#footnote-ref-5)