CA Journal of Cultural Analytics

July 8, 2020

Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

Grant Storey, David Mimno

ARTICLE INFO

Peer-Reviewed By: Folgert Karsdorp, Justin Stover

Article DOI: 10.22148/001c.13680

Dataverse DOI: 10.7910/DVN/PGHVYP

Journal ISSN: 2371-4549

ABSTRACT

One commonly recognized feature of the Ancient Greek corpus is that later texts frequently imitate and allude to model texts from earlier time periods, but analysis of this phenomenon is mostly done for specific author pairs based on close reading and highly visible instances of imitation. In this work, we use computational techniques to examine the similarity of a wide range of Ancient Greek authors, with a focus on similarity between authors writing many centuries apart. We represent texts and authors based on their usage of high-frequency words to capture author signatures rather than document topics and measure similarity using Jensen-Shannon Divergence. We then analyze author similarity across centuries, finding high similarity between specific authors and across the corpus that is not common to all languages.

Traditional analyses often assert that Ancient Greek authors imitate models set by previous authors. Arrian, writing in the second century CE, is said to have based his *Indica* on the model of Herodotus,[1] Aelius Aristides is said to have written in the style of Demosthenes and other rhetoricians,[2] and Apollonius Rhodius is said to have crafted his *Argonautica* in the style of Homer. As Antonios Rengakos puts it, the *Argonautica* is full of "imitations of Homeric phrases, verses, motifs or scenes and reproduces lexical, morphological, syntactical and metrical peculiarities of the old epic."[3] Most prior analyses focus on a few authors and examine highly visible, marked imitation, like the use of words that appear once in the entirety of the Homeric corpus (*hapax legomena*)[4] or the reuse of nearly identical phrase structures from earlier speeches.[5]

While previous studies have focused more specifically on imitation by individual authors, many Ancient Greek texts drew on prior works, with authors in genres like history, rhetoric, and poetry often writing in a specific literary dialect to match earlier works.[6] This means we might expect to see imitation of earlier models not just by a

few authors, but by many authors throughout the corpus. This hypothesis is difficult to test, as the idea of imitation involves *intention* to be similar, and intention is difficult to verify on a case-by-case basis, let alone across a corpus. In this work, we will instead focus only on similarity between authors, and leave determining author intention to other works. More specifically, we will evaluate the hypothesis that authors across the extant Ancient Greek corpus, even those writing centuries apart, are very similar to each other when compared to authors from other corpora.

To test this hypothesis, we must have a way to measure similarity between two authors or texts and baseline corpora for comparison with the Ancient Greek corpus. The similarity measurement will enable us to discuss author similarity across the corpus, and the baseline corpora will allow us to determine whether the similarity between authors in the Greek corpus is in line with patterns in any language corpus or whether it shows unusually high (or low) similarity.

Computational techniques allow us to analyze similarity in ways that go beyond focusing on a few specific instances. Recent work has begun to use computational methods to analyze classical texts, including authorship and allusion in Latin texts[7] and the syntactic style of Attic prose.[8] Our goal in this work is to analyze similarities in the writing style of Ancient Greek authors. The definition of "style" is a thorny problem that we cannot fully address within the context of this work, but at a high level we might expect the "style" of a section of text to be informed by some combination of its genre (e.g., fantasy, biography, epic poem, military history, philosophical treatise, etc.), dialect (e.g., American vs British English), time period (English as a language changes from Shakespeare to Charles Dickens to J.K. Rowling), register (simple or artistically stylized), and other internal tendencies of the author. Parts of this "style" might vary across different works from the same author, or even possibly within a text in the case of a work with multiple styles like Faulkner's *As I Lay Dying*. In this paper we focus on authors' usage of very common words, which has been used successfully in non-classics work on stylometry to capture information about an author's writing style.[9] There are a variety of potential caveats and issues which we will discuss as they come up, but the idea of an author signature provides us with a more concrete starting point for comparing different authors.

While prior works focus on using author signatures of this sort to resolve authorship questions about a given text, our work is focused on comparing word usage of *known* authors writing in different styles and time periods. There has been past work on examining imitation with regards to content words, dialect forms, or salient phrases[10] as well as analyses of allusion and intertextuality based on phrases or a small number of words.[11] However, to the best of our knowledge previous study has focused on comparing relatively short phrases between a few works, rather than considering "style" or author signature in entire texts across a larger corpus.

For our baseline comparison corpora, we choose texts from English and Icelandic. These corpora, like the Greek corpus, have a mix of genres spanning around six to ten centuries of time with texts from centuries across that span. We chose English because we expect authors to show significant differentiation over the past centuries as the language changed from "Middle" to "Modern" English.[12] Conversely, we chose Icelandic because we expect authors to show more similarity across the centuries, as linguists consider Icelandic to be a relatively conservative language in terms of change over time.[13] It is also an apt comparison for Greek because we have versions of texts from a wide range of time periods with standardized spelling and morphology. Between these two corpora – one quickly changing, one more conservative – we hope to contextualize the similarity over time in our Greek corpus. Our goal in choosing these additional corpora is to determine whether the stability we observe in Greek is universal among languages with long-term written traditions. We make no claim that the pattern we observe in Greek is unique to that language, and in fact we suspect that similar patterns may exist in other ancient languages such as Akkadian, Latin, and Sanskrit.

Using the top words as a measure of author signature and English and Icelandic as baseline corpora, we make three major contributions in this work. First, we show that a feature set based on top words captures information about authors that matches conclusions from prior non-computational scholarship in classics. Second, we explore various metrics for measuring similarity between authors and find that, for the task of predicting the work and author of text segments, a similarity metric based on Jensen-Shannon Divergence performs very well, showing 2.5% improvement over Burrow's Delta and 6% improvement over Cosine Similarity. Finally, we analyze the relationship between author similarity and relative composition date, and show that, when compared to authors in English and Icelandic, Ancient Greek

authors are far more similar across longer time periods. We find similarity between authors both at the individual author level and across the whole corpus that is significantly greater than the collections of similar chronological breadth in English and Icelandic.

Data and Corpora

Our analyses focus on texts available through the Perseus Digital Library's Greek Collection.[14] This corpus includes 464 works from 92 authors spanning from circa the 8th Century BCE to the 6th Century CE. Where possible, we divide each work into smaller segments: for example, we divide Herodotus' *Histories* (a single "work") into its nine books (each of which is a "segment"). When a work cannot be broken up naturally, it is considered a single segment, so Euripides' *Medea* is both one work and one segment. This partition leads to 1,337 total segments. Of these segments, 1,139 (from 65 authors) are prose and 198 (from 27 authors) are poetry. We remove non-Greek characters and punctuation and, where possible, restore elided tokens, so παρ' is restored to παρά. The full dataset has 9,707,987 tokens (total words) and 486,326 types (unique words) after preprocessing. See Table 1 for a top-level breakdown of the size of works and segments.

We examine the dataset in two forms. First, we group works by author, including all authors, no matter how little text there is. Each author has at least 2,000 tokens except for Bion of Phlossa (1,803 tokens) and the anonymous author of the fragmentary *Hymn to Dionysus* (just 144 tokens). The small sample size does not seem to adversely affect analysis of Bion. The short length of the *Hymn to Dionysus* does have an impact on some analyses, which we discuss below. We also analyze the texts divided into individual segments, only considering segments with at least 1,000 tokens. This gives 1,204 segments to analyze (out of the total of 1,337).

We expect that the language of Ancient Greek prose and poetry will show clear distinctions. Ancient Greek poetry, in addition to constraints of grammar, topic, and style, also had to conform to a poetic meter. There were a variety of meters used in different contexts (e.g., Dactylic Hexameter was the meter of epic poetry) but all of them mandated some pattern of long and short syllables in each line of poem. This

means that, unlike prose, all Ancient Greek poems are naturally structured around lines and have extra constraints on how sentences can be constructed.

Stylometric analysis is more uncertain for ancient documents than modern documents because of their complex chains of transmission. Unlike most work on modern authors, it is not safe to assume that the published version accurately represents the author's original work and, by extension, their author signature. Because the texts were written so long ago, our modern editions are an editor's reading of a set of medieval manuscripts, which may have long histories themselves. For example, after Euripides wrote the text of *Medea* it was passed down by actors in Athens, then standardized by scholars in Alexandria, then transmitted in a variety of manuscripts through the medieval period to the modern day, where the manuscripts were combined by an editor and then digitized to create the single version of the text which we use. Even if we find patterns in word use in modern digital texts, there is no guarantee that these patterns existed in the text as first written by Euripides. Previous work on medieval Dutch texts has even shown that copying scribes can introduce their own signature to texts.[15] Accounting for the interaction between the editorial and scribal artifacts of the manuscript tradition and the output of our method is beyond the scope of this study but is important to acknowledge.

The English corpus has 204,366,114 tokens and 701,562 types (see Table 1 for a top-level breakdown of the size of works and segments). It is a combination of the following corpora, resulting in 166 authors with 2,759 unique works and 2,960 segments:

- Modern English texts from the Gutenberg Dataset,[16] with a few duplicate texts and texts including a mixture of prose and poetry removed.
- The plays of Shakespeare from the Shakespeare Corpus.[17]
- Middle English texts from the TEAMS Middle English Text Series[18] supplemented by the Morte D'Arthur[19] and Canterbury Tales.[20]

The Icelandic corpus has 7,587,999 tokens and 290,924 types (see Table 1 for a top-level breakdown of the size of works and segments). It is a combination of the following corpora, resulting in 196 authors with 213 unique works, each of which is one segment:

- Icelandic Sagas from the Saga Corpus, with duplicate manuscripts removed.[21]
- The Icelandic Parsed Historical Corpus (IcePaHC), a collection of texts from 1150-2008 CE.[22] Duplicate manuscripts and translations were removed.
- 21st Century Icelandic texts from the Tagged Icelandic Corpus (MÍM).[23] We only used the texts labeled books, with articles by Baldur Jónsson and recipe books removed.

For both languages our goal was to create corpora with texts from a wide range of centuries by tying together multiple existing corpora from more localized time periods. This allows us to create baseline corpora that are comparable to Greek.

	5%	25%	50%	75%	95%
Greek Works	886	3,312	7,011	12,350	105,273
Greek Segments	679	2,244	4,843	9,797	21,703
English Works	4,092	22,846	57,381	96,341	192,579
English Segments	4,105	22,223	57,222	96,446	177,796
Icelandic Works	3,259	12,781	23,429	47,362	98,841

Table 1: Number of tokens in the 5th, 25th, 50th, 75th, and 95th percentile for the works and segments of different languages. While most English works and all Icelandic works consist of a single segment, we only compare these corpora at the author level, so this is not a concern for this work.

We would also like to confirm it is fair to compare usage of top words in these corpora. If the Ancient Greek vocabulary was consistent across time while the English and Icelandic vocabulary changed drastically this would provide a simple explanation for greater stability over time in Greek without the need for further exploration. In order to get a quick sanity check on change in word usage over time in our corpora, we calculate the 100 most frequent words for each century, then examine the amount of overlap in these words for time periods six centuries apart. For the English corpus, the mean and median are **48** words of overlap. For the Icelandic corpus, the mean is **68.25** and the median is **68**. For the Greek corpus, the mean is **68.14** and the median is **71**. These results show that the Icelandic corpora is a good comparison for Greek due to their similar stability of vocabulary over time. The English corpora shows less stability but seems reasonable for a comparison corpus chosen to show more change over time and fits with the claims about the relative stability of English and Icelandic.

Document Representations through High-Frequency Words

As discussed above, in order to analyze author similarity in the Ancient Greek corpus, we must have a way to measure it. To measure similarity, we first extract each author's usage of some top words as a feature vector and then use a metric to measure the similarity between these feature vectors. We begin with the process for extracting our feature vector and show that it captures useful information about authors and agrees with conclusions from prior works in classics.

In this work, our feature set for measuring author signatures is the frequency of the most frequently used tokens across the corpus. Usage of the most frequent words has shown promising results for identifying author signatures in the past, in particular because content words are more dependent on genre and topic matter.[24] Some work has found that frequent words alone can do better than part of speech information or a combination of the two.[25] The intuition behind analyzing top words is that an author's usage of these words (say, the ratio between their usage of "but" and "and") is mostly unconscious and provides a signature (or fingerprint) for that author, while their usage of content words (say, "tropical" or "lodestar") reflects the topic of their work or conscious decisions about their style rather than a fundamental characteristic of the author.

Ancient Greek is highly inflected, so a common first step when working with these texts is to replace surface forms with lemmata (e.g. "trees" to "tree" and "held" to "hold"). While choosing our top words we consider only surface forms, not lemmata, for two reasons. First, there is valuable information in the inflection of high-frequency words: for example, an author's usage of τῆς as compared to τόν is an interesting and potentially relevant distinction. Second, lemmatization is not always reliable: many tokens are ambiguous, and we are not able to resolve these across 9.7 million tokens.

When choosing our list of top words for each language, we select the 250 most frequent words across all texts that *also* occur in more than 50% of authors. This cutoff prevents inclusion of words, mostly names, that appear frequently in only a few authors. In Greek the excluded words are Ῥωμαίων, "Romans'" (27% of authors), Σωκράτης, "Socrates" (29%) and Ἀθηναῖοι, "Athenians" (50%). In

Icelandic the excluded words are *Ned* (15% of authors), *konungur*, "king" (44%) and *Jón* (45%). There are no words excluded in the English corpus. Raising the cutoff to 60% would remove a variety of non-name words including Greek ὑμᾶς, "you (pl.)," Icelandic *kannski*, "maybe," and English "should." The full list of tokens used for each language can be found in Appendix 2.

In Measuring Similarity Between Authors and Segments below we also examine a list of words including not only the top 250 words across all texts combined but also the list of the top 100 words in only the poetic texts. Due to overlap between these two sets, this yields a list of 264 total words. Common words from poetry were considered because the word usage in Ancient Greek poetry and prose has key distinctions, as we will see below. Since the corpus is dominated by prose, including more words that are specifically relevant to poetry (including more poetry-specific words like κεν and words that appear more often in poetry than prose, like Ζεύς) may help better capture the signatures of poetic authors. The full list of added words is τοι, μιν, ᾗ, ἀμφί, αὐτάρ, Διός, σ᾽, ἐνί, περ, οὔ, ἔνθα, κεν, Ζεύς, πατρός.

In the following, our feature set is the frequency of the top words within each author or segment. We therefore represent each author and segment with a vector P of 250 or 264 features, where P_i corresponds to the frequency of word i within the author or segment.

$$P_i = \frac{(\#\text{of occurrences of word } i \text{ in text } P)}{(\text{total } \#\text{of words in text } P)}$$

Note that the total number of words includes *all* words, not just the top 250/264.

Detecting Author Characteristics

Before using the set of most frequent words to analyze author similarity, we show that this feature set captures information about the texts by analyzing the Ancient Greek data to see if there are any interesting properties that can be detected from the word representation vectors.

We begin with a visualization of similarity between texts based on the top 250 words. Since the position of points in 250-dimensional space is difficult to present, we use a two-dimensional tSNE projection[26] in Figure 1. The tSNE visualization attempts

to preserve relative distances, so points that have similar word usage appear closer together. The most salient feature of Figure 1 is the large gap between authors writing prose (red) and poetry (blue), suggesting that 250 words are sufficient to distinguish these genres. In fact, the difference is so clear that the *Hymn to Dionysus* fragment is clearly recognized as poetry based on a tiny text sample (144 total tokens).

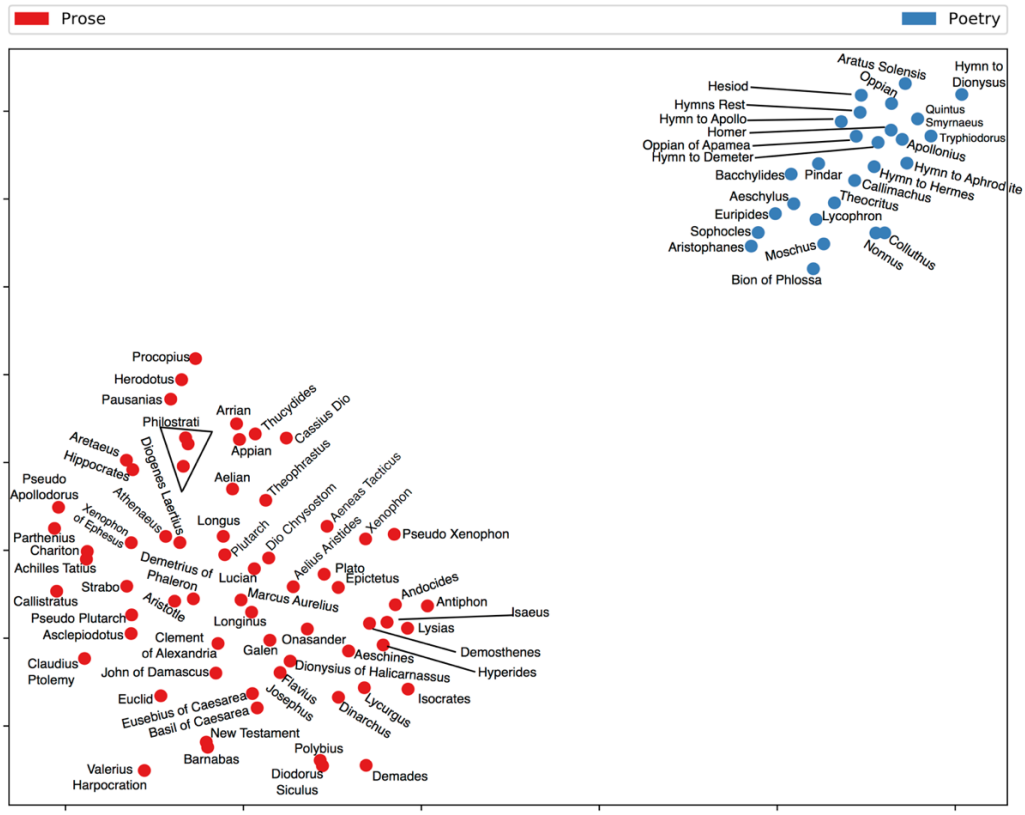

Figure 1: A two-dimensional tSNE Projection of the authors based on their usage of the top 250 words, with similar authors grouped together. Points are clustered without knowledge of their genre, so the separation between poetry and prose is based entirely on different word usage in the two genres.

The clear distinction between prose and poetry texts gives some hope that other characteristics of these authors might be distinguishable as well. In Figure 2, we show clustering based on the top 250 words, with four different colorings, based on poetry vs prose (upper left), a narrower genre distinction[27] (upper right), time period (bottom left) and dialect (bottom right). While poetry and prose form independent clusters, none of these other categories shows a clear distinction.

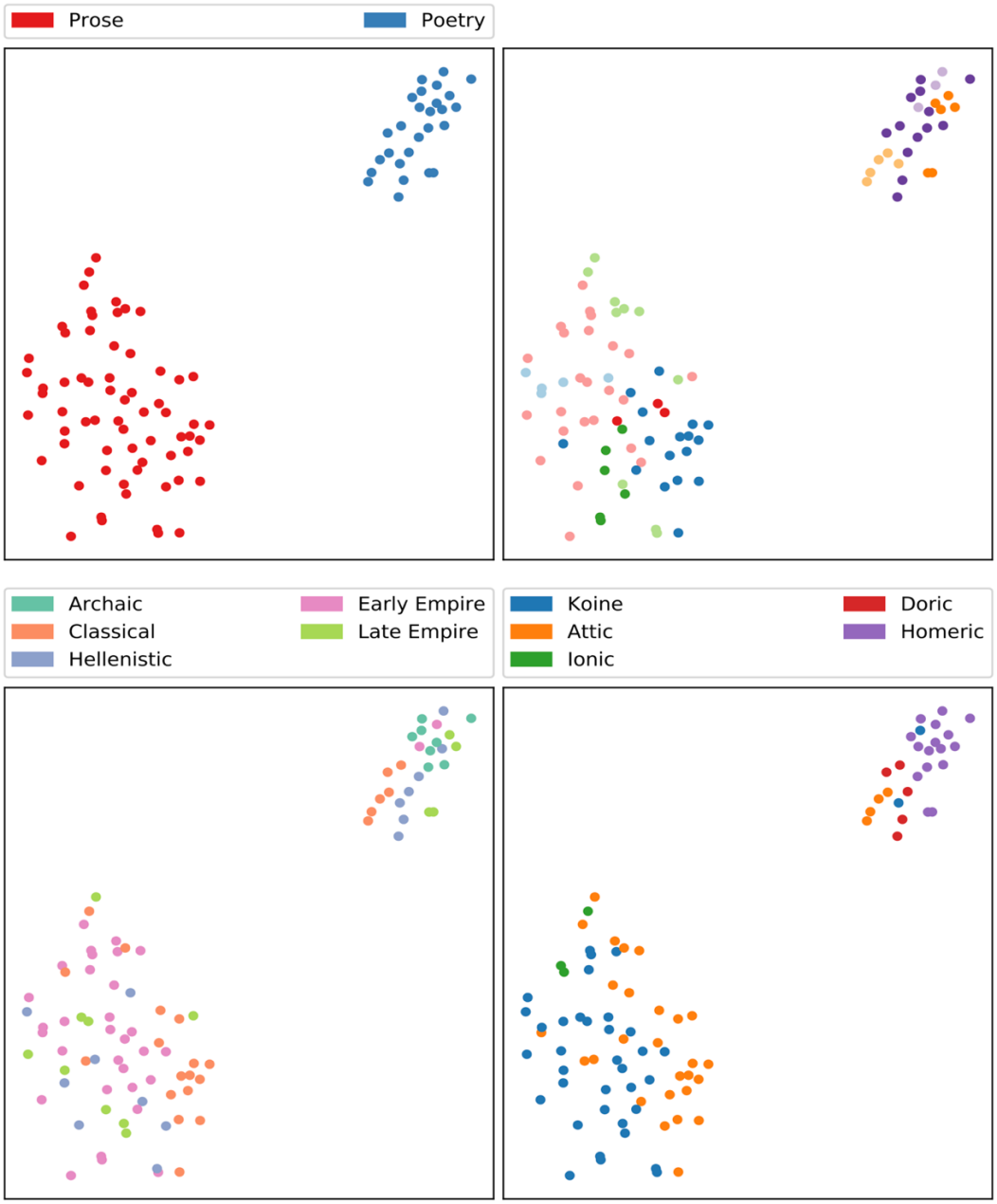

Figure 2: tSNE Projection of authors based on their usage of the top 250 words, with similar authors grouped together. The four charts show the same authors colored on different criteria: prose and poetry (top left), more narrow genres like military prose and epic poetry (top right), time period (bottom left), and dialect (bottom right).

The unsupervised tSNE Projections do not show a clear distinction between categories beyond poetry/prose, but we may still be able to detect these categories using machine learning. To test this hypothesis, we run three classifiers: A Majority Class baseline, K Nearest Neighbors (with K=2 chosen from the set {1, 2, 3, 5, 10,

20} based on best performance on the training folds) and a Multinomial Naive Bayes. For each classifier, we divide the data into nine folds. Since there are 92 authors, this provides around 10 authors to test on for each fold. The folds are chosen by randomly dividing the given author or segments into groups of roughly equal size. In addition, the segments are divided so that segments from the same work are in the training or test fold, but not both, *except* when predicting authors (marked with a *) to account for the fact that many authors in the dataset have only a single work and would therefore be impossible to accurately classify without relaxing this constraint. For each fold, we evaluate a model trained on the other eight folds. Table 2 shows the average accuracy over the nine folds for each method and each target variable.

Prediction Task	Majority Class	KNN	Naive Bayes
Genre of Authors	0.704040	**1.000000**	0.988889
Dialect of Authors	0.305051	0.729293	**0.730303**
Time Period of Authors	0.354545	**0.575758**	0.398990
Genre of Segments	0.841395	0.997512	**1.000000**
Dialect of Segments	0.282479	**0.773426**	0.723606
Time Period of Segments	0.486147	**0.658699**	0.524795
Author of Segments*	0.104696	0.858851	**0.887873**

Table 2: Results of running simple machine learning on the frequency data based on the top 250 words.

Both the KNN and Naive Bayes classifiers do extremely well at predicting genre (poetry vs prose) of authors and segments, achieving >98% accuracy in all cases, which is in line with prior state of the art techniques.[28] Dialect and time period prediction for authors are slightly worse, but still far better than the majority class baseline. When considering individual segments, K Nearest Neighbors performs better than the majority class at predicting dialect and time period. Because the test set does not include segments from books in the training set, this accuracy is not based on detecting segments from the same book. Allowing segments from the same book in both folds increases the accuracy of KNN to 96% and 95% for dialect and time period prediction respectively.

The final row of Table 2 also shows the accuracy of these classifiers at predicting the author of a segment. Naive Bayes accurately classifies 88% of segments, with KNN performing almost as well. This performance is not perfect and could certainly be improved, but it does show that the features are at least partially predictive of authorship, even with relatively simple techniques.

When considering the authors together, the data can be used to predict dialect and time period much better than a simple baseline, but not well enough to confidently classify every text. With the larger amount of data present in the segment-by-segment analysis, we can predict dialect, time period, and to a lesser extent author with high accuracy. At the author level, dialect and time period are perhaps hard to predict with the small amount of data, but at the segment level they are reasonably predictive, even if they do not show clearly distinct clusters.

These results could perhaps be improved by using more complex classifiers with greater hyperparameter tuning, but that task is beyond the scope of this paper. Even these results show that the feature set we have chosen — examining the frequency of the top words within an individual segment or an author's work as a whole — captures information about these texts including genre (poetry vs prose), and, when there are many samples in the segment case, dialect, time period, and authorship.

Analyzing Author Segments

We have seen that our feature set captures relevant information about our texts; we now show it provides results that agree with prior work in classics. One frequent area of discussion in classics is the authorship of specific texts. When considering works or segments attributed to a specific author, we expect that segments which are outliers based on our feature set were either written by a different author or written in a very different style. To visualize this, we create a 2-dimensional tSNE projection of the many different segments, as we did for the authors, and analyze segments that are outliers compared to the author's other segments. For the most part, segments by similar authors cluster together nicely, but there are some exceptions. Each panel in Figure 3 highlights a single author's segments in blue, showing one clear outlier.

Figure 3: Four charts showing segments that are outliers for different authors. Segments written by the author highlighted in each subplot are blue, while all remaining segments from the Ancient Greek corpus are gray. Segments are grouped using a two-dimensional tSNE projection.

In the upper left, we see speeches attributed to Demosthenes, with speech 59 (*Against Neaera*) distinct from the rest. Critics since Dionysius of Halicarnassus have considered this speech to be by an author other than Demosthenes,[29] with some modern scholars attributing the text to Apollodorus.[30] On the other hand, the upper

right shows that Isocrates' Speech 21, *Against Euthynus*, is also a clear outlier, but the attribution of this speech to Isocrates is quite secure.[31] While the text is almost certainly the work of Isocrates, it *is* recognized as being in a markedly different style.

The other two cases are more complex. Xenophon's authorship of the *Cynegeticus* (bottom left) has long been both challenged and defended, but it is unquestionably written in a very different style from the rest of Xenophon's work.[32] Our analysis (bottom right) supports a recent study by Thomas Koentges that found Plato's *Menexenus* to be the most unusual of his works based on a few computational analyses.[33] This text has been suspected of being non-Platonic in the past, but not recently.[34] Although our analysis highlights the *existence* of differences, further work would be required to distinguish the "different author" and "same author, unusual style" possibilities.

Across a few different data analyses, we have shown that high-frequency words capture valuable information about authors and texts, including genre, time period, dialect, authorship, and style. Based on the performance of our simple classifier with respect to various author characteristics and the agreement with prior work concerning outlier segments from certain authors, we can be confident that these features are capturing characteristics of the authors and will be a good basis for comparing varied writing signatures.

Measuring Similarity Between Authors and Segments

Given our feature set, the second part of calculating similarity is a using consistent metric for comparing top word use from different authors. We use the Jensen-Shannon Divergence metric, finding that it slightly outperforms other similarity metrics at predicting the author and work of individual segments.

A variety of methods have been used in authorship attribution, including comparing raw frequencies,[35] using Euclidean distance,[36] and bootstrap consensus trees,[37] among others. One of the most commonly used metrics is Burrows' Delta, which compares normalized relative usage of top words.[38] Zhao et al. found Kullback-Leibler Divergence (KL-Divergence) to be a useful metric for determining authorship of a text,[39] but it is not symmetric, meaning that in general $KL(A, B) \neq$

KL(B, A). We therefore consider Jensen-Shannon Divergence, a symmetric version of KL-Divergence. Jensen-Shannon Divergence is:

$$\text{JS Divergence}(P, Q) = \frac{1}{2}\left(\text{KL}\left(P, \frac{P+Q}{2}\right) + \text{KL}\left(Q, \frac{P+Q}{2}\right)\right)$$

Where KL is Kullback-Leibler Divergence,

$$\text{KL}(P, Q) = \sum_i P_i \ln \frac{P_i}{Q_i}$$

To get a similarity metric rather than a divergence, we calculate

$$\text{JS Similarity}(P, Q) = 1 - \text{JS Divergence}(P, Q)$$

Since this metric compares two probability distributions, we include the total number of non-top words in addition to each of the top words. We calculate probabilities using the frequency of words in each author, work, or segment, and to prevent probabilities of zero we add 1 to the count of each word before computing this frequency. The metric is symmetric and decomposes over individual words, meaning that we can tell how much each individual word (καί, δέ, etc.) contributed to the similarity between two authors. Cosine Similarity, which is commonly used for comparison, does not have this feature.

Metric Evaluation

To evaluate the performance of Jensen-Shannon Similarity, we compare its performance to four other common similarity metrics: Burrows' Delta, discussed above, a popular metric for authorship analysis,[40] Manhattan Distance (which focuses on absolute differences in word usage), Canberra Distance (which focuses on relative differences in word usage), and Cosine Distance.

We expect that for a given segment, the most similar segment should be from the same work and by the same author. To test this, we first examine all segments that have another segment from the same work (e.g., *Iliad* book 1 has all the other books of the *Iliad* in the same work) and determine whether the most similar segment is from the same work using a leave-one-out validation method. This validation is equivalent to a K Nearest Neighbors classifier with 1 neighbor and distance

determined by the given metric. We next compare all segments that have another segment by the same *author* and determine whether the most similar segment is by the same author, again using leave-one-out.

For each metric, we also examine its performance when using the top 250 overall words and the top 250 overall words + top 100 poetry words (for a total of 264 words). As a comparison for the significance of our metrics, we consider two baselines: Cosine Similarity, which is commonly used for comparing distributions, and Burrows' Delta, which is used for analyzing authorship based on top words. For both metrics, we choose the top 250 overall words as our baseline to determine whether using extra poetry words shows improvement.

Metric	Top 250	Top 250 + Top 100 in Poetry
Jensen-Shannon	**92.40**%†‡	92.28%†‡
Burrows' Delta	88.95%	89.55%†
Manhattan	88.00%	88.24%
Canberra	88.00%	88.24%
Cosine	86.46%	86.82%

Table 3: How well similarity metrics based on a given set of words identify whether two segments come from the same work. †: Results very significant ($p < 0.01$) when compared to Cosine (250). ‡: Results very significant ($p < 0.01$) when compared to Burrows' Delta (250).

Metric	Top 250	Top 250 + Top 100 in Poetry
Jensen-Shannon	**92.20**%†‡	92.03%†‡
Burrows' Delta	89.83%†	89.92%†
Manhattan	89.92%†	90.00%†
Canberra	88.47%	88.31%
Cosine	87.63%	87.97%

Table 4: How well similarity metrics based on a given set of words identify whether two segments come from the same author. †: Results very significant ($p < 0.01$) when compared to Cosine (250). ‡: Results very significant ($p < 0.01$) when compared to Burrows' Delta (250).

When comparing segments from the same work (Table 3), Jensen-Shannon Similarity based on the top 250 words (M=0.924, SD=0.265) does the best job of identifying segments, significantly better than both Cosine Similarity (M=0.865, SD=0.342), t(841)= -6.617, p=6.523e-11 and Burrows' Delta (M=0.890, SD=0.313), t(841)=-3.944, p=8.681e-05. When comparing segments from the same author (Table 4), Jensen-Shannon Similarity based on the top 250 words (M=0.922,

SD=0.268) also does the best job of identifying segments and again shows very significant improvement over both Cosine Similarity (M=0.876, SD=0.329), t(1179)=-6.130, p=1.195e-09 as well as Burrows' Delta (M=0.898, SD=0.302), t(1179)=-3.517, p=4.534e-04. In both cases, Jensen-Shannon Similarity performs the best of the metrics with high significance ($p < 0.01$). This advantage is not due to plus-one smoothing or the number of unrecorded non-top words: adding this information does not significantly improve the other methods. We also recognize that because Jensen-Shannon Similarity relies on probability estimates, it may suffer from bias when working with smaller texts. Even with this bias, it shows strong performance on these prediction tasks, including analysis of all of the shortest Ancient Greek segments. Considering this, we proceed with the similarity metric while keeping an eye out for patterns that appear to result from biased probability estimates from small segments rather than actual signal in the text.

It is also worth noting that adding 14 extra poetry words does not always improve accuracy, and for Jensen-Shannon it actually decreases it. Given this result, we will use the top 250 words **without** poetry words for further analysis, which allows us to avoid potential bias from using a metric constructed with some knowledge of the corpus (i.e. poetry vs prose authors).

These results show that Jensen-Shannon Similarity with the top 250 words is the best metric at identifying work or author, and we can be confident that high similarities according to this metric are based on actual similarities of the texts. We also note that Gerlach and Font-Clos independently found Jensen-Shannon Divergence to be a useful metric for comparing texts of different genres.[41] It is possible that there may be better metrics available, or performing preprocessing such as document normalization may improve the performance of some or all of our metrics. Our goal, however, is to find a metric that performs well enough at detecting similarity to be useful for further analyses of similarity, not to find the absolute best possible metric, and we believe Jensen-Shannon Similarity meets this standard.

Analyzing High-Similarity Authors

Now that we have established that our feature set and metric form a good measurement of similarity between different segments, we can begin evaluating

whether authors from our Ancient Greek corpus are more similar across long time periods than authors from our English and Icelandic corpora. We begin by taking the top 100 closest author pairs in Ancient Greek and looking at the pairs of authors writing at least four centuries apart. In the Ancient Greek corpus, there are 23 such pairs, which is far more than we find in the baseline corpora.

Of these top 100 closest author pairs in Ancient Greek, 23 pairs wrote at least four centuries apart. These pairs fall into a few clear categories:

Epic Poets: This category consists of epic poets spanning from Homer to the late Roman Empire: Apollonius similar to Homer; Apollonius, Oppian, and Oppian of Apamea similar to Hesiod; Oppian, Oppian of Apamea, and Tryphiodorus similar to Apollonius; Quintus Smyrnaeus similar to both Apollonius and Homer.

Attic Style: This category consists of authors in Imperial Rome writing with signatures like those of orators and prose authors from the golden age of Athens: Aelius Aristides and Dio Chrysostom similar to Aeschines, Demosthenes, and Plato; Aelius similar to Andocides; Dio Chrysostom similar or Xenophon; and Appian and Arrian similar to Thucydides.

Christian Authors: This category consists of John of Damascus (c. 700 CE) writing with a signature similar to three prior Christian authors: Clement of Alexandria, Eusebius of Caesarea, and Basil of Caesarea.

Eusebius and Dionysius: Eusebius of Caesarea (c. 300 CE) and Dionysius of Halicarnassus (first century BCE). Eusebius is closer to signature of the Jewish author Flavius Josephus (the 16th most similar pair in our dataset), who wrote about a century after Dionysius with a similar signature (these two are the 8th most similar pair in our set). The result indicates that Eusebius writes like Flavius Josephus, who in turn writes like Dionysius of Halicarnassus. This is similar to the Attic orators above: Aelius Aristides did not attempt to imitate every one of these authors at once, but wrote in an Attic style based at times on Demosthenes, Isocrates, Plato, and Xenophon.[42] Aelius therefore appears similar to, say, Aeschines due to his similarity to Demosthenes and Isaeus, who are in turn very similar to Aeschines.

These authors are more similar than authors writing at the same time in similar genres, such as Isocrates and Lysias (members of the "Ten Attic Orators"), Plato and Xenophon, or Aratus and Callimachus, and they are far more similar than the

historians Herodotus and Thucydides (697th most similar). So there does seem to be evidence of high similarity between authors writing in similar genres, even centuries apart.

Authors writing at least four centuries apart but still showing high levels of similarity also appear more frequently in the Ancient Greek corpus than the baselines. Of the top 100 (2.4%) pairs of Greek authors, 23 (23%) wrote at least four centuries apart. Of the top 2.4% pairs of Icelandic authors, 2.6% (12/457) are writing at least four centuries apart, and of the top 2.4% pairs of English authors, 0% (0/328) are writing at least four centuries apart; in fact, none of these top pairs are writing more than one century apart. So Ancient Greek shows a much higher occurrence of temporally distant authors in the most similar pairs. There is no plausible case that the similarity between these author pairs is due simply to random variance present in all languages; as the English corpus makes clear, it is not even a necessary feature of a corpus.

These similar authors are also reasonably consistent across the other metrics considered, so when compared using other metrics in Table 5 the pairs usually appear in the top 100 pairs and in all but six cases appear in the top 10% (418) of pairs. The six pairs not in the top 10% according to Canberra distance are later epic poets compared to earlier ones, which is likely due to Canberra distance focusing too much on the relative frequencies of less frequent words. This shows that the similarity seen is not just an artifact of our Jensen-Shannon similarity metric but reflects underlying similarity between these authors picked up by other metrics like Burrow's Delta, Manhattan Similarity, and Cosine Similarity.

Authors	Jensen-Shannon	Burrow's Delta		Manhattan		Canberra		Cosine	
Apollonius/Quintus Smyrnaeus	10	11	(-1)	55	(-45)	114	(-104)	21	(-11)
Apollonius/Homer	12	8	(4)	29	(-17)	417	(-405)	4	(8)
Apollonius/Oppian	19	35	(-16)	120	(-101)	221	(-202)	47	(-28)
Apollonius/Oppian of Apamea	23	66	(-43)	227	(-204)	**745**	**(-722)**	132	(-109)
Aelius Aristides/Demosthenes	25	23	(2)	9	(16)	5	(20)	18	(7)
Homer/Quintus Smyrnaeus	26	16	(10)	142	(-116)	**453**	**(-427)**	36	(-10)
Clement/John of Damascus	40	40	(0)	57	(-17)	88	(-48)	45	(-5)
Hesiod/Oppian of Apamea	42	81	(-39)	115	(-73)	**624**	**(-582)**	10	(32)
Aelius Aristides/Plato	43	74	(-31)	27	(16)	7	(36)	51	(-8)
Dionysius/Eusebius	44	58	(-14)	40	(4)	33	(11)	84	(-40)
Eusebius/John of Damascus	59	64	(-5)	86	(-27)	105	(-46)	54	(5)

Apollonius/Hesiod	60	59	(1)	302	(-242)	**626**	**(-566)**	212	(-152)
Aeschines/Aelius Aristides	63	99	(-36)	53	(10)	41	(22)	74	(-11)
Hesiod/Oppian	69	105	(-36)	324	(-255)	**831**	**(-762)**	70	(-1)
Apollonius/Tryphiodorus	70	185	(-115)	437	(-367)	**1166**	**(-1096)**	27	(43)
Basil/John of Damascus	71	55	(16)	77	(-6)	37	(34)	376	(-305)
Appian/Thucydides	74	46	(28)	50	(24)	69	(5)	34	(40)
Arrian/Thucydides	80	87	(-7)	24	(56)	178	(-98)	13	(67)
Demosthenes/Dio Chrysostom	85	148	(-63)	52	(33)	20	(65)	109	(-24)
Andocides/Aelius Aristides	93	190	(-97)	89	(4)	32	(61)	187	(-94)
Aeschines/Dio Chrysostom	95	242	(-147)	94	(1)	95	(0)	218	(-123)
Dio Chrysostom/Plato	96	200	(-104)	95	(1)	25	(71)	116	(-20)
Dio Chrysostom/Xenophon	98	135	(-37)	107	(-9)	28	(70)	262	(-164)

Table 5: Rank of highly similar authors writing at least four centuries apart by different metrics. Numbers in parentheses indicate difference from Jensen-Shannon rank.

Composition Date and Author Similarity

The previous section focused on a few key similar authors, but we now turn to answering our original question for the whole corpus: is there evidence that authors in the Ancient Greek corpus are more similar across time than authors from the English and Icelandic corpora? To evaluate this for a given corpus we plot each pair of authors as a single point, with the centuries between those authors on the x-axis and the similarity between those authors on the y-axis. By running a linear regression on this dataset, we can get a numerical value for the correlation between distance in time and similarity, i.e. a measurement of how similar authors are across time. We can then examine the measurements for each corpus and compare them. We find that temporal distance between authors explains only 3% of the variance in the Greek dataset but 58% and 39% of the variance in the English and Icelandic corpora, respectively – that is, Greek authors writing far apart in time are far more similar than authors from the English and Icelandic baselines.

When analyzing the Ancient Greek corpus, we make a few adjustments. First, we remove the *Hymn to Dionysus* and Euclid – the *Hymn to Dionysus* has only 144 tokens, and the texts of Euclid are full of geometric figures, so between the two of them they account for nearly every low-similarity outlier. Second, we remove author pairs more than nine centuries apart. Third, we color each of the points based on whether the authors are writing in the same genre or different genres; as seen in

Figure 4, this helps explain a bi-modal feature of the data where authors writing in the same genre are generally more similar than authors writing in different genres.

The Greek results are visible in Figure 4. There is a downward slope in similarity as authors write further apart in time, but this is mostly the result of relative frequency of prose and poetry authors over time, and only explains 3% of the variance (R2=0.03474, F(3842)=138.3, p=2.190e-31). When considering only authors writing in the same genre, or only authors writing in a different genre, the slopes are 47% and 14% as steep as the overall slope. For different genres, the century explains 0.17% of the variance and we cannot reject the hypothesis that the slope is flat (R2=1.663e-03, F(1559)=2.597, p=0.1072), despite the large number of points. When we consider the trend line for texts of the same genre, only 2.6% of the variance is explained by the century (R2=0.02592, F(2281)=60.708, p=9.973e-15). While distance in time explains very little of the variance, when the whole corpus is considered, the matchup of genre (same genre vs different genre) accounts for almost 66% of the variation seen (R2=0.6570, F(3842)= 7360.029, p<2.0E-307).

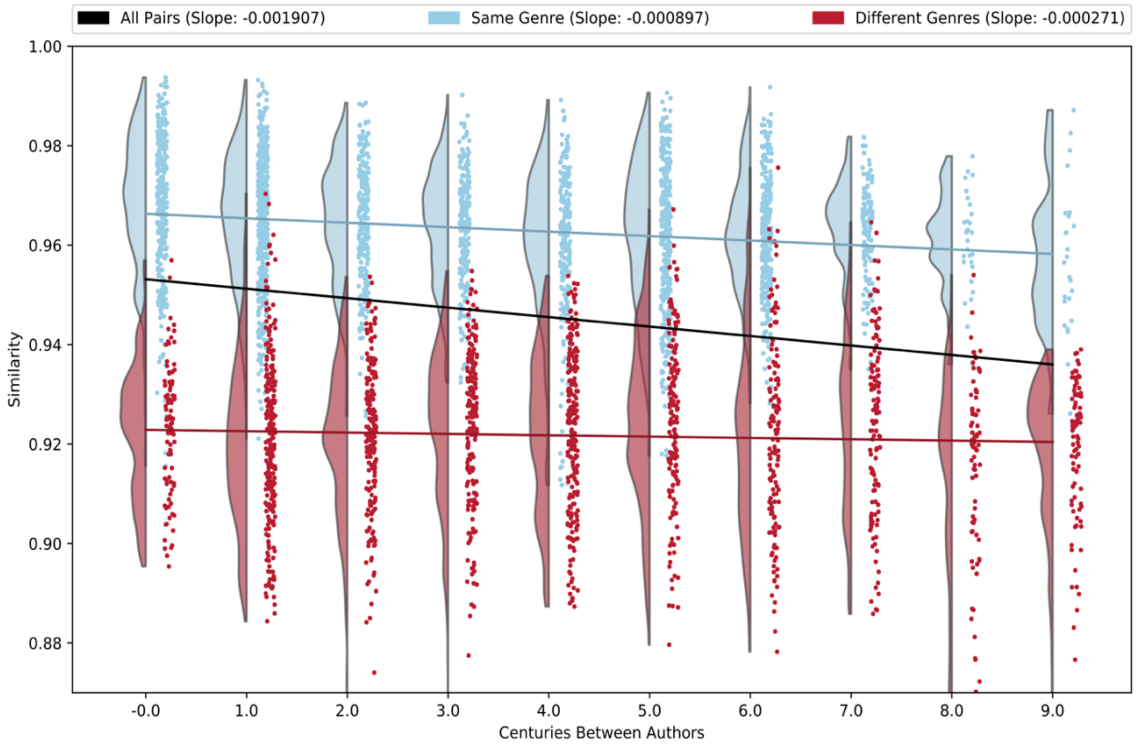

Figure 4: Graph of the similarity of pairs of Ancient Greek authors across different times, with author pairs writing in the same genre marked separately from those writing in different genres. Each dot represents a pair of authors, and the half-violin plots show the density of points.

Figure 5 shows the same analysis for pairs of English authors. Authors writing further apart in time are more different, and the slope for authors of the same genre is around 12 times steeper than for Greek: -0.0111 vs -0.0009. In addition, the century explains 58% of the variance in the texts (R2=0.5821, F(13693)= 19075.874, p<2.0E-307), while, contrary to the Ancient Greek, the genre matchup explains roughly 11% (R2=0.1071, F(13693)= 1642.735, p<2.0E-307).

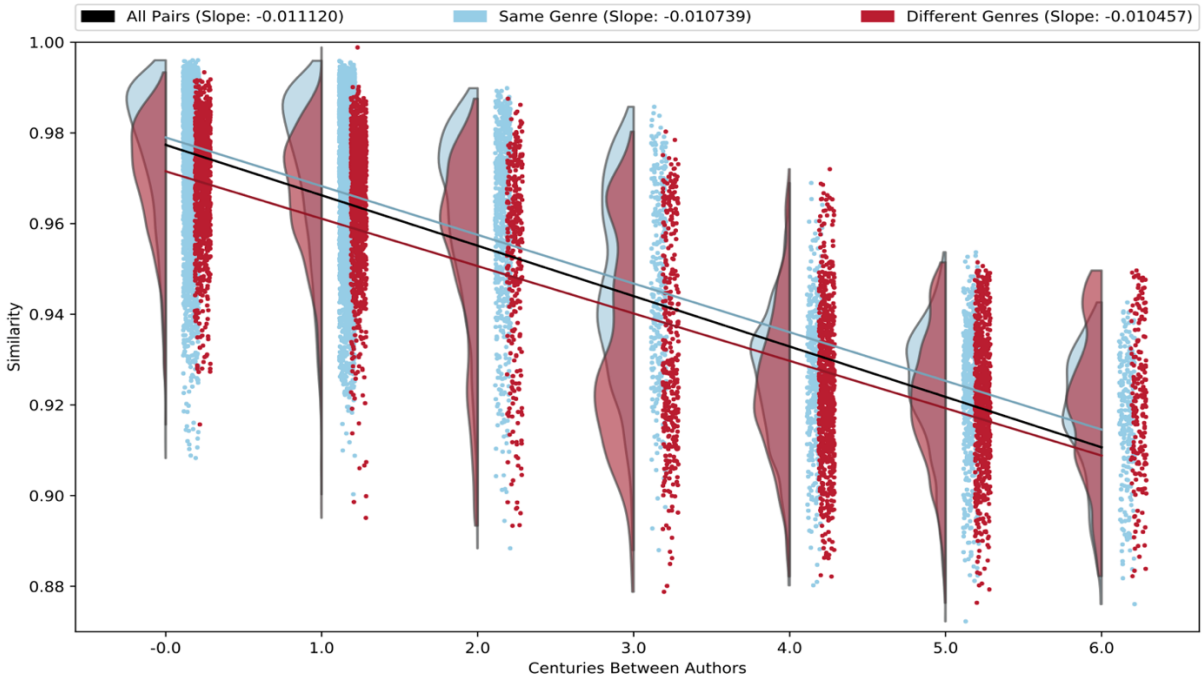

Figure 5: Graph of the similarity of pairs of Modern and Middle English authors across different times, with author pairs writing in the same genre marked separately from those writing in different genres. Each dot represents a pair of authors, and the half-violin plots show the density of points.

Figure 6 shows Icelandic author pairs. Like English, as the temporal distance between authors increases there is a clear decrease in similarity, though it accounts for only 39% of the variance rather than 58% (R2=0.3883, F(19108)= 12127.527, p<2.0E-307). There are a few pairs four to seven centuries apart that are more similar than the bulk of authors for that time, because three 19th and 20th century authors wrote sagas that are similar to older models, but there is still a clear downward trend over the corpus as a whole.

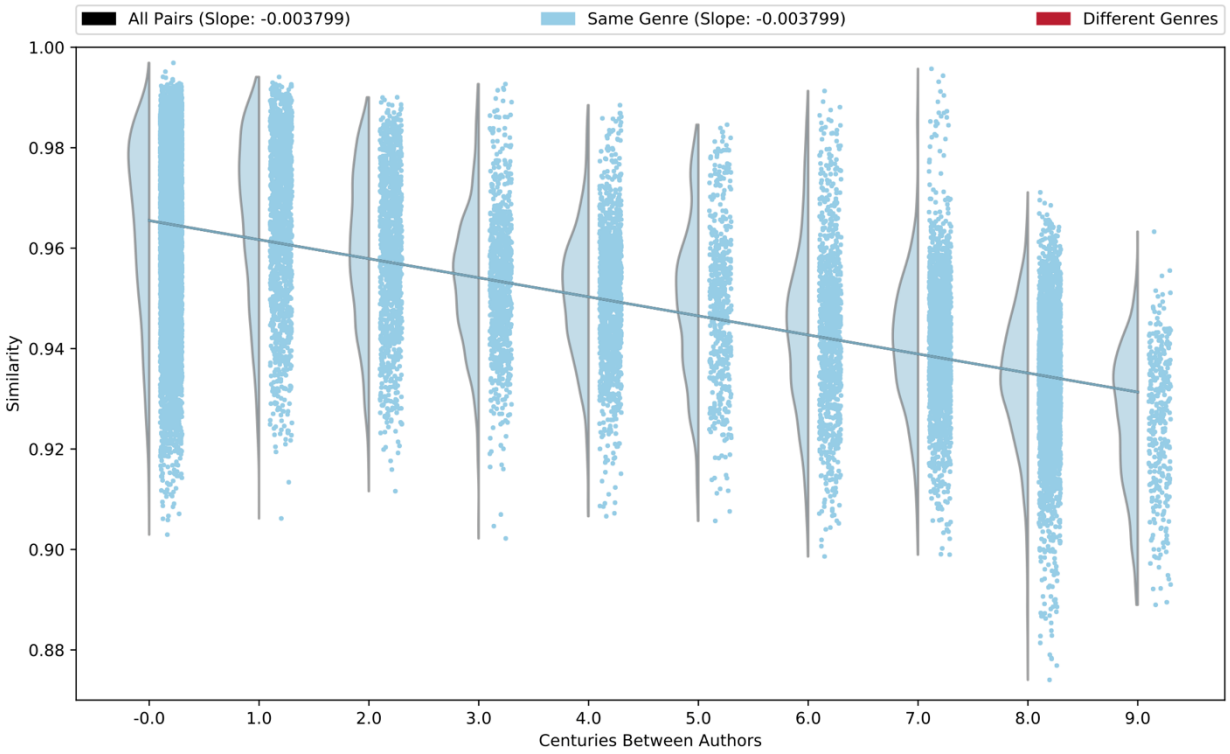

Figure 6: Graph of the similarity of pairs of Icelandic authors across different times. All texts are prose, so there are no instances of authors of different genres. Each dot represents a pair of authors, and the half-violin plots show the density of points.

The English and Icelandic data supports the hypothesis that English has undergone more change than Icelandic in the recent past.[43] However, the Greek texts show even *less* change than Icelandic over time, and far more instances of high similarity across a long time period. While we saw that century difference explains only 3% of the variance in the Ancient Greek Corpus, it explains 39-58% for English and Icelandic. The Ancient Greek Corpus shows remarkable stability even when compared to a corpus from Icelandic, a language that is culturally conservative and morphologically and orthographically standardized across time; with this context it is clear that the observed similarity between authors writing many centuries apart is a significant feature of Ancient Greek and not a given for any language. These results provide strong support for our hypothesis: it does appear that authors in the extant Ancient Greek corpus are more similar across time than authors in our baseline comparison corpora.

Conclusion

We now consider *why* Ancient Greek author similarity is so remarkably strong across time, in stark contrast to English and Icelandic where distance in time correlates more closely with difference between texts. The similarity between authors across long time periods was *not* due to a similarity in the spoken language across those time periods. Many of the later authors would have spoken the Koine (common) dialect of Greek rather than the Attic, Ionic, or Homeric dialects of the texts they imitated. When Callimachus wrote in the Homeric dialect, which combines features of Ionic and Aeolic, both dialects had been mostly replaced by the Attic-based Koine, and by the time of Aelius Aristides and his contemporaries "those who wanted to write the best Attic... clearly needed help... no one had spoken the prescribed model Attic for centuries."[44]

Genre has a significant influence on the similarity of Ancient Greek texts: prose works from six centuries apart are usually more similar than a prose-poetry pair written in the same century. Because works of poetry had additional constraints compared to prose, they also had different word usage: for example, the constraints of meter made it more difficult to use the article with every noun. The observed similarity across time is likely due in part to the strength and consistency of these poetic constraints.

Another likely source of similarity is the connection between genre and dialect. In the classical period different genres were generally associated with specific dialects, so authors wrote lyric poetry in the associated dialect, literary Doric, regardless of their own native dialect.[45] Later authors, especially orators and epic poets, seem to have followed this model and intentionally written in older dialects. It is even possible that this similarity is *caused* by the fact that the later authors were writing in a non-native dialect of the language. If Aelius Aristides spoke Koine Greek but wanted his speeches to sound like Demosthenes, he had to learn Attic Greek *from* earlier authors like Demosthenes; there were no native Attic speakers to teach the language. Since he would therefore be learning Demosthenian specifically rather than Attic, it makes sense that his speeches may look more similar to Demosthenes than a generic Attic text. There is in fact evidence from ancient authors that memorizing entire texts by heart and imitating prior works were seen as ways of

developing and improving one's own style.[46] It seems reasonable that memorization and imitation of an earlier author may lead to a writing style similar to that author.

This similarity may also have been driven by culture. Similarity to older models was the culturally correct way to write epic poems, speeches, high-register histories, and so on. Authors would have had a cultural incentive to do a good job at this, since a work that looked like the older models would be more well-regarded. There could also be survivor bias in the corpus: the works that survive to today are, for the most part, the best works by the best authors, according to cultural standards that viewed writing in certain styles as "the best." If the only Greek speeches from the Roman Empire preserved to the modern day are the ones that are "the best," where the best is defined as the most like speeches from Classical Athens, this could provide an explanation for the strength of this effect across the corpus.

These similarity results provide support for many existing claims of imitation, including the cases of Arrian, Aelius Aristides, and Apollonius discussed in the introduction. They may also provide a starting point for examinations of imitation by other less commonly treated authors in the Greek corpus, and with comparison to further corpora could be used to evaluate claims about the presence of imitation throughout the corpus. Similarity and potential imitation by Greek authors many centuries apart could also provide an interesting case study for ideas about imitation and language from cultural evolution, including language change and the value of multiple teachers.[47] Imitation is discussed as a tool for improving writing and style in the modern day, so an examination of the practice and impact of this phenomenon in more ancient texts would provide an interesting comparison.[48]

Though we see these similarities to an unusual degree in the Ancient Greek corpus, this is likely not a capability that only Ancient Greek-speaking humans had. Modern research has shown that author signatures are not immutable: there are examples of authors varying their own signature in different works or even within a single book.[49] However, there is less modern work on how an author might adjust their signature to be more like a specific model: the ancient sources suggest copying and memorization would help, but this hypothesis has not been proven.[50] One future direction of exploration is examining mechanisms for achieving this similarity: we find no evidence that authors are copying segments of text, but we could not establish whether they are copying sentence templates. Another interesting question would be

whether modern individuals could achieve this sort of similarity as well. For example, would actors who have memorized plays of Shakespeare be able to mimic the word frequencies of Shakespeare more accurately than a control group? In addition, comparison to further languages may yield additional interesting results. There are many more languages with multi-century literary traditions and the potential for imitation that could be compared to the Ancient Greek system and help clarify how much of an outlier this tradition is.

While this work suggests a variety of further areas of exploration, it shows that Ancient Greek authors of the Hellenistic and Roman periods wrote in a remarkably similar fashion to their predecessors in the classical period, at least based on their usage of common words. Further, comparisons to English and Icelandic show that this is not a natural feature of every language. Even though we approached this problem from a different direction than the usual Classics approach, we hope this, too, is instructive. As is increasingly being recognized, computational techniques can supplement and work alongside more traditional methods in Classics scholarship, providing useful context, answering questions in different ways, and opening new doors for further study of classical texts.

Acknowledgements

Many thanks to Professor Jeffrey Rusten for his invaluable guidance with the classical portions of this paper. We would also like to thank our reviewers for their helpful feedback on an earlier version of this work. This work was supported by NSF #1652536, and the Alfred P. Sloan Foundation.

Appendix 1: Code

The code for this work, along with instructions for acquiring the corpora used, is available to view or download at https://github.com/twopis/twopis.

This work uses the packages *scipy*,[51] *numpy*,[52] *scikitlearn*,[53] and *statsmodels*[54] for data processing and analysis and *matplotlib*[55] and RainCloud plots[56] for charts.

Appendix 2: Top Words

Tables 6, 7, and 8 show the top 250 words in Greek, English, and Icelandic respectively.

Rank	Token	Rank	Token	Rank	Token	Rank	Token
1	καί	**64**	νῦν	**127**	ὅπως	**190**	πολλούς
2	δέ	**65**	οὐδέν	**128**	πρότερον	**191**	ὄντα
3	τῶν	**66**	ἐξ	**129**	αὐτό	**192**	πλῆθος
4	τήν	**67**	ὧν	**130**	ὅταν	**193**	τινά
5	τό	**68**	ὥστε	**131**	τινα	**194**	πολλοί
6	μέν	**69**	αὐτός	**132**	πολύ	**195**	που
7	τοῦ	**70**	ὅ	**133**	ἡμᾶς	**196**	ἐνταῦθα
8	τῆς	**71**	ἐγώ	**134**	γενέσθαι	**197**	ποιεῖν
9	τόν	**72**	μοι	**135**	ἡμῶν	**198**	αὐτῇ
10	ἐν	**73**	ὥσπερ	**136**	αὐτῆς	**199**	πάντας
11	γάρ	**74**	πάντα	**137**	ἀεί	**200**	τούτους
12	τε	**75**	αὐτούς	**138**	ἦν	**201**	δύναμιν
13	ὁ	**76**	ἐστιν	**139**	λόγον	**202**	τρόπον
14	τά	**77**	οὕτως	**140**	πόλεως	**203**	αὐτά
15	τούς	**78**	ἄρα	**141**	λέγειν	**204**	καλῶς
16	τοῖς	**79**	μᾶλλον	**142**	τούτῳ	**205**	αὖ
17	πρός	**80**	ὑπέρ	**143**	ταύτην	**206**	εἶπεν
18	ἐπί	**81**	αἱ	**144**	ὅν	**207**	ὅσον
19	τῷ	**82**	ἔφη	**145**	μηδέν	**208**	ἕν
20	οἱ	**83**	ἤδη	**146**	ὥς	**209**	ἀρχῆς
21	ὡς	**84**	ἐπεί	**147**	εἰπεῖν	**210**	ἔπειτα
22	ἀλλά	**85**	οὕτω	**148**	ὑμᾶς	**211**	μέγα
23	ἤ	**86**	μάλιστα	**149**	οὗ	**212**	ἐκείνου
24	κατά	**87**	τότε	**150**	ἔχων	**213**	ἐγένετο
25	εἰς	**88**	ἐστι	**151**	τίς	**214**	ἥ
26	μή	**89**	πόλιν	**152**	οἷον	**215**	θεοῦ
27	ἄν	**90**	ἡμῖν	**153**	μέντοι	**216**	ἀνδρῶν
28	περί	**91**	οὐχ	**154**	ὧν	**217**	τοιαῦτα
29	οὐ	**92**	τ᾽	**155**	σε	**218**	μέρος
30	τῇ	**93**	μόνον	**156**	εἴη	**219**	ἑαυτοῦ
31	ἡ	**94**	πολλά	**157**	σύν	**220**	γῆν
32	τάς	**95**	πρῶτον	**158**	ἔστιν	**221**	πρίν
33	διά	**96**	δεῖ	**159**	χρόνον	**222**	ἄλλους
34	οὐκ	**97**	τούτου	**160**	εὖ	**223**	γῆς
35	ἐκ	**98**	ἐστίν	**161**	πάντες	**224**	ἡμεῖς

36	ὅτι	**99**	ἅμα	**162**	ἀνθρώπων	**225**	ἀνήρ
37	ἐς	**100**	μηδέ	**163**	ᾧ	**226**	βασιλεύς
38	ὑπό	**101**	αὐτήν	**164**	αὐτοί	**227**	λόγος
39	οὖν	**102**	μήτε	**165**	ἄλλα	**228**	χώραν
40	εἶναι	**103**	ἄ	**166**	μέχρι	**229**	πόλεμον
41	εἰ	**104**	ἵνα	**167**	ὅσα	**230**	βασιλέως
42	παρά	**105**	με	**168**	τοίνυν	**231**	ταύτῃ
43	ταῦτα	**106**	τούτοις	**169**	ὑμῶν	**232**	οὐδείς
44	δή	**107**	σύ	**170**	ἄνδρες	**233**	πολλάκις
45	τοῦτο	**108**	οἵ	**171**	εὐθύς	**234**	ὁμοίως
46	ἀπό	**109**	ἄλλων	**172**	ἐπειδή	**235**	ἄλλοις
47	οὐδέ	**110**	ὅς	**173**	δοκεῖ	**236**	ταύτης
48	μετά	**111**	πάντων	**174**	ἦσαν	**237**	ἔσται
49	αὐτόν	**112**	μήν	**175**	χρή	**238**	καθάπερ
50	τι	**113**	πάλιν	**176**	ἄλλο	**239**	πόλεις
51	ἦν	**114**	ἐστί	**177**	ἐμοί	**240**	αὖθις
52	γε	**115**	σοι	**178**	ποτε	**241**	πᾶν
53	αὐτῷ	**116**	ἔστι	**179**	ἤν	**242**	εἶτα
54	αὐτοῦ	**117**	οὗτος	**180**	πρό	**243**	λέγω
55	τις	**118**	ὑμῖν	**181**	πάνυ	**244**	ᾗ
56	αὐτῶν	**119**	ἔχει	**182**	οὕς	**245**	οἶμαι
57	οὔτε	**120**	ἐάν	**183**	ἀρχήν	**246**	πλήν
58	ὦ	**121**	εἴ	**184**	ὕστερον	**247**	ἄλλοι
59	ταῖς	**122**	πῶς	**185**	θεῶν	**248**	λόγῳ
60	τούτων	**123**	δύο	**186**	πολλῶν	**249**	ἄνδρα
61	αὐτοῖς	**124**	τοῦτον	**187**	οὗτοι	**250**	τινες
62	τί	**125**	οἷς	**188**	πόλει		
63	ἔτι	**126**	ἔχειν	**189**	ὅτε		

Table 6: List of tokens used and their rank in the top 250 tokens found in all Greek texts (excluding tokens appearing in 50% of authors or fewer).

Rank	**Token**	**Rank**	**Token**	**Rank**	**Token**	**Rank**	**Token**
1	the	**64**	man	**127**	don't	**190**	quite
2	and	**65**	into	**128**	life	**191**	asked
3	of	**66**	some	**129**	being	**192**	night
4	to	**67**	little	**130**	away	**193**	because
5	a	**68**	has	**131**	thought	**194**	father
6	in	**69**	than	**132**	still	**195**	moment
7	i	**70**	about	**133**	through	**196**	work
8	that	**71**	mr	**134**	went	**197**	heard
9	he	**72**	like	**135**	though	**198**	few
10	was	**73**	time	**136**	yet	**199**	knew

11	it	**74**	upon	**137**	just	**200**	told
12	his	**75**	any	**138**	without	**201**	enough
13	as	**76**	did	**139**	last	**202**	between
14	you	**77**	our	**140**	while	**203**	course
15	with	**78**	only	**141**	take	**204**	find
16	for	**79**	other	**142**	many	**205**	love
17	had	**80**	see	**143**	nothing	**206**	part
18	is	**81**	know	**144**	mrs	**207**	side
19	not	**82**	should	**145**	hand	**208**	seen
20	but	**83**	can	**146**	young	**209**	years
21	her	**84**	well	**147**	sir	**210**	also
22	at	**85**	before	**148**	every	**211**	each
23	be	**86**	down	**149**	eyes	**212**	miss
24	on	**87**	such	**150**	once	**213**	among
25	she	**88**	much	**151**	get	**214**	both
26	have	**89**	two	**152**	off	**215**	lord
27	him	**90**	after	**153**	place	**216**	perhaps
28	by	**91**	made	**154**	face	**217**	whole
29	they	**92**	its	**155**	ever	**218**	having
30	which	**93**	us	**156**	found	**219**	heart
31	this	**94**	these	**157**	people	**220**	whom
32	all	**95**	must	**158**	let	**221**	round
33	my	**96**	may	**159**	same	**222**	it's
34	from	**97**	great	**160**	another	**223**	god
35	said	**98**	come	**161**	tell	**224**	poor
36	so	**99**	good	**162**	house	**225**	almost
37	were	**100**	over	**163**	under	**226**	however
38	me	**101**	how	**164**	things	**227**	home
39	we	**102**	here	**165**	right	**228**	want
40	or	**103**	old	**166**	look	**229**	yes
41	one	**104**	never	**167**	head	**230**	room
42	there	**105**	go	**168**	why	**231**	hands
43	no	**106**	say	**169**	got	**232**	soon
44	are	**107**	think	**170**	left	**233**	indeed
45	if	**108**	own	**171**	looked	**234**	woman
46	would	**109**	men	**172**	thing	**235**	door
47	their	**110**	first	**173**	saw	**236**	oh
48	an	**111**	way	**174**	mind	**237**	name
49	been	**112**	came	**175**	put	**238**	myself
50	when	**113**	himself	**176**	always	**239**	turned
51	them	**114**	where	**177**	seemed	**240**	rather
52	what	**115**	might	**178**	give	**241**	end

53	who	**116**	am	**179**	three	**242**	called
54	will	**117**	again	**180**	lady	**243**	felt
55	out	**118**	back	**181**	something	**244**	nor
56	up	**119**	too	**182**	going	**245**	anything
57	do	**120**	those	**183**	against	**246**	mother
58	more	**121**	even	**184**	done	**247**	dear
59	then	**122**	make	**185**	better	**248**	since
60	now	**123**	long	**186**	world	**249**	matter
61	could	**124**	day	**187**	far	**250**	country
62	your	**125**	shall	**188**	took		
63	very	**126**	most	**189**	new		

Table 7: List of tokens used and their rank in the top 250 tokens found in all English texts.

Rank	**Token**	**Rank**	**Token**	**Rank**	**Token**	**Rank**	**Token**
1	og	**64**	sé	**127**	gekk	**190**	farið
2	að	**65**	mig	**128**	þessa	**191**	hægt
3	í	**66**	síðan	**129**	kemur	**192**	láta
4	á	**67**	hafi	**130**	komið	**193**	hét
5	er	**68**	hér	**131**	vildi	**194**	sínu
6	sem	**69**	henni	**132**	hins	**195**	fer
7	hann	**70**	fór	**133**	eitthvað	**196**	hélt
8	til	**71**	þér	**134**	öðrum	**197**	ganga
9	var	**72**	vel	**135**	spurði	**198**	öll
10	en	**73**	sá	**136**	hinn	**199**	fóru
11	við	**74**	inn	**137**	sama	**200**	íslandi
12	það	**75**	hefði	**138**	þannig	**201**	kannski
13	um	**76**	þær	**139**	verða	**202**	stundum
14	ég	**77**	aftur	**140**	getur	**203**	áfram
15	ekki	**78**	hennar	**141**	manna	**204**	gerði
16	með	**79**	varð	**142**	allir	**205**	tekið
17	þá	**80**	áður	**143**	okkar	**206**	fannst
18	af	**81**	hjá	**144**	árið	**207**	minn
19	því	**82**	maður	**145**	stað	**208**	meðal
20	fyrir	**83**	fara	**146**	fá	**209**	mundi
21	hún	**84**	saman	**147**	eitt	**210**	það
22	þeir	**85**	hana	**148**	fyrr	**211**	átt
23	þar	**86**	undir	**149**	sagt	**212**	orð
24	svo	**87**	heldur	**150**	hver	**213**	öllu
25	sér	**88**	tók	**151**	þann	**214**	svona
26	eftir	**89**	sínum	**152**	annars	**215**	of
27	þegar	**90**	þessu	**153**	bæði	**216**	þótti
28	hafði	**91**	þessi	**154**	sú	**217**	dag

29	upp	**92**	koma	**155**	nema	**218**	yrði
30	frá	**93**	segja	**156**	öllum	**219**	undan
31	honum	**94**	höfðu	**157**	lét	**220**	skal
32	voru	**95**	gera	**158**	oft	**221**	sögu
33	þetta	**96**	vegna	**159**	gert	**222**	já
34	þess	**97**	mjög	**160**	meira	**223**	gott
35	mér	**98**	okkur	**161**	fyrst	**224**	finna
36	eða	**99**	átti	**162**	fékk	**225**	kvað
37	þeim	**100**	hvort	**163**	stóð	**226**	væru
38	eru	**101**	enn	**164**	orðið	**227**	ár
39	verið	**102**	má	**165**	sitt	**228**	gæti
40	hans	**103**	leið	**166**	komu	**229**	samt
41	nú	**104**	ekkert	**167**	þig	**230**	máli
42	hafa	**105**	sinni	**168**	einu	**231**	veit
43	úr	**106**	bara	**169**	fólk	**232**	enginn
44	eins	**107**	mun	**170**	halda	**233**	mönnum
45	sagði	**108**	niður	**171**	síðar	**234**	miklu
46	þú	**109**	verður	**172**	sín	**235**	vissi
47	út	**110**	sinn	**173**	meðan	**236**	hátt
48	þeirra	**111**	mikið	**174**	án	**237**	nær
49	sig	**112**	sína	**175**	alltaf	**238**	landi
50	væri	**113**	milli	**176**	alla	**239**	ráð
51	vera	**114**	einn	**177**	vér	**240**	þarna
52	hefur	**115**	taka	**178**	vegar	**241**	mál
53	þau	**116**	heim	**179**	aðeins	**242**	vill
54	ef	**117**	aldrei	**180**	rétt	**243**	lengi
55	fram	**118**	þótt	**181**	enda	**244**	séu
56	hvað	**119**	annað	**182**	geta	**245**	vita
57	allt	**120**	þessum	**183**	skyldi	**246**	þarf
58	eigi	**121**	tíma	**184**	ofan	**247**	margir
59	kom	**122**	hvernig	**185**	mín	**248**	jafnvel
60	þó	**123**	mælti	**186**	eiga	**249**	nokkuð
61	menn	**124**	líka	**187**	utan	**250**	innan
62	yfir	**125**	einnig	**188**	gat		
63	segir	**126**	sjá	**189**	móti		

Table 8: List of tokens used and their rank in the top 250 tokens found in all Icelandic texts (excluding tokens appearing in 50% of authors or fewer).

Notes

[1] Hammond, N. G. L. "The speeches in Arrian's Indica and Anabasis." *The Classical Quarterly* 49, no. 1 (1999): 238-253.

[2] Jones, Christopher P. "Aelius Aristides, ΕΙΣ ΒΑΣΙΛΕΑ." *The Journal of Roman Studies* 62 (1972): 134-152.

[3] Rengakos, Antonios. "Apollonius Rhodius as a Homeric Scholar." In *Greek Literature in the Hellenistic Period*, pp. 241-264. Routledge, 2018.
[4] Kumpf, Michael Martin. "The Homeric hapax legomena and their literary use by later authors, especially Euripides and Apollonius Rhodius." PhD diss., The Ohio State University, 1975.; and Appel, Włodzimierz. "Die homerischen hapax legomena bei Quintus Smyrnaeus: Adverbien." *Glotta* 71, no. 3./4. H (1993): 178-188.

[5] Jones 1972: 134-152.

[6] Tribulato, Olga. "Literary dialects." *A companion to the ancient Greek language* (2010): 388.

[7] Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. "Authenticating the writings of Julius Caesar." *Expert Systems with Applications* 63 (2016): 86-96.; and Stover, Justin, and Mike Kestemont. "Reassessing the Apuleian corpus: a computational approach to authenticity." *The Classical Quarterly* 66, no. 2 (2016): 645-672.

[8] Bozia, Eleni. "Atticism: The Language of 5th-century Oratory or a Quantifiable Stylistic Phenomenon?." *Open Linguistics* 1, no. open-issue (2016).; and Bozia, Eleni. "Measuring Tradition, Imitation, and Simplicity: The case of Attic Oratory." *Corpus-Based Research in the Humanities (CRH)* (2015): 23.

[9] Burrows, John. "'Delta': a measure of stylistic difference and a guide to likely authorship." *Literary and linguistic computing* 17, no. 3 (2002): 267-287.

[10] Schmid, Wilhelm. *Der Atticismus in seinen Hauptvertretern von Dionysius von Halikarnass bis auf den zweiten Philostratus*. Vol. 4. W. Kohlhammer, 1896; and Clark, Donald Lemen. "Imitation: Theory and practice in Roman rhetoric." *Quarterly Journal of Speech* 37, no. 1 (1951): 11-22.

[11]Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. "The Tesserae Project: intertextual analysis of Latin poetry." *Literary and linguistic computing* 28, no. 2 (2012): 221-228.; and Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson. "Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching." *Digital Scholarship in the Humanities* 30, no. 4 (2015): 503-515.

[12]Milroy, James, and Lesley Milroy. "Linguistic change, social network and speaker innovation." *Journal of linguistics* 21, no. 2 (1985): 339-384.; and Friðriksson, Finnur. *Language change vs. stability in conservative language communities. A case study of Icelandic*. Institutionen för lingvistik, 2008.

[13] Milroy and Milroy; Friðriksson; Árnason, Kristján. *The phonology of Icelandic and Faroese*. Oxford University Press, 2011.

[14] Smith, David A., Jeffrey A. Rydberg-Cox, and Gregory R. Crane. "The Perseus Project: A digital library for the humanities." *Literary and Linguistic Computing* 15, no. 1 (2000): 15-25.

[15] Kestemont, Mike, and Karina van Dalen-Oskam. "Predicting the past: memory based copyist and author discrimination in Medieval epics." In *Proceedings of the Twenty-first Benelux Conference on Artificial Intelligence (BNAIC 2009)*, pp. 121-128. Eindhoven: Benelux Association for Artificial Intelligence, 2009.; and van Dalen-

Oskam, Karina, and Joris Van Zundert. "Delta for middle Dutch—author and copyist distinction in Walewein." *Literary and Linguistic Computing* 22, no. 3 (2007): 345-362.

[16] Lahiri, Shibamouli. "Complexity of Word Collocation Networks: A Preliminary Structural Analysis." In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 96-105. 2014.

[17] Scott, Mike. "Shakespeare Corpus." *https://lexically.net/wordsmith/support/shakespeare.html*.

[18] Peck, Russell, ed. "TEAMS Middle English Text Series." *http://d.lib.rochester.edu/teams/text-online*.

[19] Malory, Thomas. *Le Morte Darthur*. University of Michigan Humanities Text Initiative, 1997. http://name.umdl.umich.edu/MaloryWks2.

[20] Chaucer, Geoffrey. *Chaucer's Works, Volume 4 (of 7) — The Canterbury Tales*. University of Oxford, 2007. *https://www.gutenberg.org/files/22120/22120-h/22120-h.htm*.

[21] Rögnvaldsson, Eiríkur, and Sigrún Helgadóttir. "Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change." In *Language Technology for Cultural Heritage*, pp. 63-76. Springer, Berlin, Heidelberg, 2011.

[22] Joel C. Wallenberg, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. "Icelandic Parsed Historical Corpus Version 0.9." 2011. *http://www.linguist.is/icelandic_treebank*.

[23] Helgadóttir, Sigrún, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir, and Hrafn Loftsson. "The tagged Icelandic corpus (MÍM)." *Language Technology for Normalisation of Less-Resourced Languages* (2012): 67.

[24] Burrows 2002.

[25] van Dalen-Oskam, Karina. "The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant's Scolastica (1271)." *Literary and linguistic computing* 27, no. 4 (2012): 355-372.

[26] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9, no. Nov (2008): 2579-2605.

[27] between Romance, Speeches, Military/Historical Prose, Judeo-Christian Prose, Philosophy, Other Prose, Comedy/Tragedy, Epic Poetry, Didactic Poetry, and Other Poetry.

[28] See Gianitsos, Efthimios, Thomas Bolt, Pramit Chaudhuri, and Joseph Dexter. "Stylometric Classification of Ancient Greek Literary Texts by Genre." In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 52-60. 2019. *The Greek corpus used in this work is slightly different from ours, preventing direct comparison of results.*

[29] Dionysius. *The Critical Essays*. Harvard University Press, 2015. *https://www.loebclassics.com/view/dionysius_halicarnassus-style_demosthenes/1974/pb_LCL465.239.xml*.

[30] Kapparis, Konstantinos A. Apollodoros "Against Neaira"[D 59]. Vol. 53. Walter de Gruyter, 2012.

[31] Isocrates, and Larue Van Hook. *Isocrates*. Harvard University Press, 2015. *https://www.loebclassics.com/view/isocrates-discourses_21_euthynus/1945/pb_LCL373.351.xml*.

[32] Gray, Vivienne J. "Xenophon's' Cynegeticus'." *Hermes* 113, no. H. 2 (1985): 156-172.

[33] Koentges, Thomas. "Computational Analysis of the Corpus Platonicum." Tech. rep., Center for Hellenic Studies, Harvard University, 2018.

[34] Tsitsiridis, Stavros. *Platons Menexenos: einleitung, text und kommentar*. Vol. 107. Walter de Gruyter, 2011.

[35] Binongo, José Nilo G. "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution." *Chance* 16, no. 2 (2003): 9-17.

[36] Labbé, Dominique. "Experiments on authorship attribution by intertextual distance in English." *Journal of Quantitative Linguistics* 14, no. 1 (2007): 33-80.

[37] van Dalen-Oskam, Karina. "Epistolary voices. the case of Elisabeth Wolff and Agatha Deken." *Literary and Linguistic Computing* 29, no. 3 (2014): 443-451.

[38] Burrows, 2002; Þorgeirsson, Haukur. "How similar are Heimskringla and Egils saga? An application of Burrows' delta to Icelandic texts." *European Journal of Scandinavian Studies* 48, no. 1 (2018): 1-18.; Barnes-Sadler, Simon. "A Digital Humanities Approach to Inter-Korean Linguistic Divergence: Stylometric Analysis of ROK and DPRK Journalistic Texts." *S/N Korean Humanities* 4, no. 1 (2018): 127-153.; Benatti, Francesca, and Justin Tonra. "English Bards and Unknown Reviewers: a Stylometric Analysis of Thomas Moore and the Christabel Review." *Breac: A Digital Journal of Irish Studies* (2015).; Eder, Maciej. "Style-markers in authorship attribution: a cross-language study of the authorial fingerprint." *Studies in Polish Linguistics* 6, no. 1 (2011).; and van Dalen-Oskam and Zundert.

[39] Zhao, Ying, Justin Zobel, and Phil Vines. "Using relative entropy for authorship attribution." In *Asia Information Retrieval Symposium*, pp. 92-105. Springer, Berlin, Heidelberg, 2006.

[40] Burrows, 2002.

[41] Gerlach, Martin, and Francesc Font-Clos. "A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics." *Entropy* 22, no. 1 (2020): 126.

[42] Jones 1972.

[43] Milroy and Milroy.

[44] Horrocks, Geoffrey. *Greek: A History of the Language and its Speakers*. John Wiley & Sons, 2014.

[45] Tribulato.

[46] Lucian. *How to Write History*. Harvard University Press, 1959. https://www.loebclassics.com/view/lucian-how_write_history/1959/pb_LCL430.3.xml.; and Clark, 12-13.; and Corbett, Edward PJ. "The Theory and Practice of Imitation in Classical Rhetoric". *College Composition and Communication* (1971): 243-250.

[47] Mesoudi, A. *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press. 2011.; Tamariz, Monica. "Experimental studies on the cultural evolution of language." *Annual Review of Linguistics* 3 (2017): 389-407.; and Loke, James Wai Chuen. "Are Five Teachers Better Than One? The Effect of Multiple Models on Cultural Transmission." *Undergraduate Journal of Psychology* (2013): 47.

[48] D'Angelo, Frank J. "Imitation and style." *College Composition and Communication* 24, no. 3 (1973): 283-290.; Geist, Uwe. "Stylistic imitation as a tool in writing pedagogy." In *Effective Learning and Teaching of Writing*, pp. 169-179. Springer, Dordrecht, 2005..

[49] Evans, Mel. "Tudor women writing: Multimodal style and identity in the English letters and prose of Queen Katherine Parr and Princess Elizabeth." *Studies in Variation, Contacts and Change in English* 17 (2016).; Pearl, Lisa, Kristine Lu, and Anousheh Haghighi. "The character in the letter: Epistolary attribution in Samuel Richardson's Clarissa." *Digital Scholarship in the Humanities* 32, no. 2 (2017): 355-376.; and Burrows, John. "Who wrote Shamela? Verifying the authorship of a parodic text." *Literary and linguistic computing* 20, no. 4 (2005): 437-450.

[50] Clark, 12-13.

[51] Jones, Eric, Travis Oliphant, and Pearu Peterson. "SciPy: Open source scientific tools for Python." (2001).

[52] van der Walt, Stéfan, S. Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation." *Computing in Science & Engineering* 13, no. 2 (2011): 22-30.

[53] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

[54] Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python." In *Proceedings of the 9th Python in Science Conference*, vol. 57, p. 61. Scipy, 2010.
[55] Hunter, John D. "Matplotlib: A 2D graphics environment." *Computing in science & engineering* 9, no. 3 (2007): 90-95.

[56] Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. “Raincloud plots: a multi-platform tool for robust data visualization.” *Wellcome open research*, 4 (2019).