Author identification of short texts using dependency treebanks without vocabulary

Robert Gorman  
Classics & Religious Studies, University of Nebraska-Lincoln,  
Lincoln, Nebraska, USA

Abstract

How to classify short texts effectively remains an important question in computational stylometry. This study presents the results of an experiment involving authorship attribution of ancient Greek texts. These texts were chosen to explore the effectiveness of digital methods as a supplement to the author’s work on text classification based on traditional stylometry. Here it is crucial to avoid confounding effects of shared topic, etc. Therefore, this study attempts to identify authorship using only morpho-syntactic data without regard to specific vocabulary items. The data are taken from the dependency annotations published in the Ancient Greek and Latin Dependency Treebank. The independent variables for classification are combinations generated from the dependency label and the morphology of each word in the corpus and its dependency parent. To avoid the effects of the combinatorial explosion, only the most frequent combinations are retained as input features. The authorship classification (with thirteen classes) is done with standard algorithms—logistic regression and support vector classification. During classification, the corpus is partitioned into increasingly smaller ‘texts’. To explore and control for the possible confounding effects of, e.g. different genre and annotator, three corpora were tested: a mixed corpus of several genres of both prose and verse, a corpus of prose including oratory, history, and essay, and a corpus restricted to narrative history. Results are surprisingly good as compared to those previously published. Accuracy for fifty-word inputs is 84.2–89.6%. Thus, this approach may prove an important addition to the prevailing methods for small text classification.

1 Introduction

How to improve accuracy when identifying authorship in short texts is an open question in computational stylometry. Some work has shown promising results in classification tests involving a small number of author classes. Hirst and Feiguina (2007) discriminate 200-word passages of Charlotte or Anne Brontë with a best accuracy of 92.4%. Gamon (2004) includes writings by Emily Brontë as well for a three class problem. His best accuracy for twenty-sentence chunks (roughly 500–600 words) is 96–98%, for five-sentence chunks (roughly 100–120 words) it is 85%.

While such accuracy with classes of two or three authors may be helpful with some real-world problems, many open questions of authorial identity involve a larger number of possible answers where
classification is more difficult. The issue has recently attracted scholarly attention: Luyckx and Daelemans (2008, 2011) explore the negative correlation between number of authors and accuracy, and Eder (2015) has systematically investigated small text size in multiclass categorization and has suggested that, speaking generally, 5,000 words is a minimum for useful results. Later, considering the issue more closely, Eder has found that 2,000 words may be sufficient when the texts in question ‘exhibit a clear authorial signal’ (Eder, 2017). Of course, even the smaller of these thresholds is much too large for many authorship problems. This study explores syntactic features drawn from dependency treebanks as the possible basis for author identification in significantly smaller texts.

Naturally enough, most work on machine text classification has focused on modern languages, with English leading the way. The present author, however, is a classicist whose research involves evaluating the authorship of short passages in collections of ancient Greek historical texts (Gorman and Gorman, 2016). Thus, the material used here is drawn from a corpus in that language. In addition, most attempts at computational authorship attribution depend to a greater or lesser degree on measures of vocabulary richness. With this approach, the particular words chosen by a writer are taken to be constitutive of a stylometric ‘authorial signature’ allowing texts to be distinguished. On the other hand, the ‘content words’ of a text are highly sensitive to factors besides authorship: genre, topic, addressee, and so on. For this reason, some researchers prefer an admixture of ‘function words’ (e.g. prepositions, articles, and conjunctions) and/or syntactic features to make analyses less subject to such confounding influences. However, in a recent review of authorship classification, Stamatakos (2016, p. 9) notes that ‘it is not yet possible to extract stylometric measures that are only determined by the personal style of author and are immune to changes in topic or genre’. This study will try to improve upon this situation by moving further away from measures of content words. It is particularly important to avoid any possible confounding effects of subject matter, because the short Greek texts that are the author’s ultimate concern are found in collections arranged by topic, with a large proportion of shared vocabulary. Any method that fails to take account of this lexical overlap will be found unconvincing. Thus, the present investigation avoids completely any consideration of vocabulary richness and pays no notice to the occurrence of individual words. Instead, syntactic features, representing both shallow and deep linguistic characteristics, make up the only data to be analyzed.

The rest of this article is organized as follows. First, an initial corpus of poetry and prose input texts is identified and the information available through its annotations is explained. The next two sections detail how the corpus annotations were used to generate a range of input features and then how the huge number of possible features was very severely culled to constitute a feasible set of independent variables. After this, the design of the classification test procedure is outlined and the results presented. Next, possible confounding elements in the mixed corpus of poetry and prose are identified. After this, the results of classifying more focused corpora, first of prose texts alone, then of only narrative histories, are presented. Finally, the results are examined more closely, and a possible reason for their success is suggested.

2 The Corpus

The texts for our experiments come from the Ancient Greek and Latin Dependency Treebank (AGLDT, Bamman and Crane, 2011), published online by the Perseus Digital Library and the affiliated Perseids Project.¹ The AGLDT is an excellent source of data for investigations involving syntactic features, because the texts are hand annotated and of generally high quality. For ancient Greek, the collection includes texts in various genres of poetry and prose, with the date of composition ranging across a period of roughly 1,000 years. The corpus selected for the first experiment reported here contains twenty-eight texts by thirteen different authors.² These works consist of 582,487 tokens in total. The corpus is quite heterogeneous, with several distinct genres of poetry and prose. Admittedly,
using such a mélange for a classification test runs the risk of reintroducing the possible confounding effects of genre and topic, and adding that of the difference between annotators. However, when this experiment was being designed and prepared, it seemed necessary to ignore these considerations in order to assemble the largest possible test corpus. More recently, newly available texts have enabled us to confront these confounders directly, as we will see below.

Each text is represented by an XML (eXtensible Markup Language) file detailing lexical, morphological, and syntactic information for each token:

```xml
<word id="1" form="Θούκυδίδης" lemma ="Θούκυδίδης" postag="n-s—mn-" relation ="SBJ" head="3"/>
<word id="2" form="Ἀθηναῖος" lemma ="Ἀθηναῖος" postag="n-s—mn-" relation ="ATR" head="1"/>
<word id="3" form="γυνέραμψε" lemma ="συγγράφω" postag="v3siaia—" relation ="PRED" head="0"/>
<word id="4" form="τον" lemma="τ" postag="l-s—ma-" relation ="ATR" head="5"/>
<word id="5" form="πόλεμου" lemma ="πόλεμος" postag="n-s—ma-" relation ="OBJ_AP" head="10"/>
```

This is the dependency annotation for the first words of Thucydides’ Histories: ‘Thucydides of Athens composed an account of the war’. According to the AGLDT annotation scheme, each word element contains an index number (the ‘id’ attribute), the token (‘form’), the lexical type (‘lemma’), a morphological description (‘postag’), a label for the type of the syntactic dependency (‘relation’), and an index number for the syntactic dependency parent of the current word (‘head’).

3 The Features

The data contained in the AGLDT offer an important advantage for the creation of features upon which to base a syntax-driven attribution experiment. They record information both about what the attribution literature refers to as ‘shallow’ syntactic characteristics as well as ‘deep’ analytic features. The shallow features are essentially those noted in the ‘postag’ attribute of the annotation: the morphological description of a given word. A word’s part of speech or its tense or gender, for example, are syntactic data points, but they do not allow us to identify the syntactic structure of a sentence. On the other hand, the ‘relation’ attribute is clearly structural: a word is identified as subject, adverb, etc., according to its function in the sentence structure. Less obvious, perhaps, is the way in which the ‘head’ attribute of each word can be used to generate deep syntactic information from morphological data. The ‘head’ attribute gives the index of the parent of the target word. With this information, each word in a sentence can be easily associated with any of its ancestors or descendants, and with any features of those words that may be recorded in the annotation. In other words, the ‘head’ attribute allows information about the structure of any portion of the sentence dependency tree to be included.

The AGLDT data allow the structure of the whole-sentence dependency tree or any of its subtrees to be used to generate input features. In this study, dependency information for two ‘generations’ is admitted: the morphological characteristics and the dependency relation for each word were supplemented with the equivalent data for its syntactic parent. This method lends a deeper syntactic dimension to much data that would otherwise be ‘merely’ morphological, since any feature that encodes a dependency relationship ipso facto reflects syntactic structure. For example, a word encoded morphologically as a ‘noun singular feminine accusative’ is simultaneously grouped with its parent, a ‘verb present indicative active 3rd person singular’. Thus, morphological combinations that are generated by syntactic relationships are introduced into the data to produce a richer representation of the input text. At the same time, with regard to including syntactic structural features in authorship classification tests, more is not necessarily better. Here, all data for more than one generation of dependency ‘ancestors’, as well as data for all ‘sibling’ words, have been excluded. This restriction helps to reduce the negative effects of the
exponential increase in data (and sparseness) entailed in examining connected groups of words.  

4 Feature Selection

The morphological annotation of each word in the AGLDT gives data on nine distinct categories. In addition, all words also have a single annotation for dependency relationship. Thus, there are twenty categories of morpho-syntactic data available for each word, when the target item and its parent are considered together. Combinations of these categories form the basis for our classification experiments. However, before going on to extract the data for each word, it is necessary to reduce sharply the number of possible combinations: if we include all combinations from sets of one category (e.g. only grammatical number or only dependency relation) to the maximal set of all twenty categories, there will essentially be \(2^{20}\) possible combinations. These combinations represent ‘feature types’ whose number, as we will see, must eventually be multiplied by a corresponding number of ‘feature values’. Let us take as an example the first word of the annotation of Thucydides given above: assuming a feature type which combines the dependency relation of the target word with that of its parent, Word 1 of Thucydides would be assigned a combined value of ‘SBJ’ and ‘PRED’. There are of course many possible values for each feature type and their relative frequencies are the independent variables in our attribution tests.

With combinations of twenty syntacto-grammatical categories yielding more than a million feature types, and with each type serving as the basis for the generation of a group of different values, it is clear that the number is far too large to be practical and some selection must be made. At the same time, it is not possible to foresee which feature types are likely to be most valuable for successful classification. Therefore, we have chosen rather arbitrary cut-offs to select a practical number of features for consideration. First, since sparsity increases rapidly with the number of elements combined in a feature type, we limit the complexity of these types to combinations of no more than five categories; the individual elements of each combination are drawn, without replacement, from any of the twenty available categories (i.e. \(\binom{20}{1}\) \ldots \(\binom{20}{5}\)). This step cuts the number of feature types to something less than 22,000—still far too many. On the assumption that the most frequent feature types will be the most useful for classification, the next step was to rank selected types by frequency. To count the occurrences of a particular feature type, it is necessary to populate it with a value for each associated token. Because this process is slow, even on a relatively fast computer, the full corpus was not used in this step. A smaller sample made up of 1,000 tokens randomly drawn from each text was used for feature selection.

Feature types are evaluated for inclusion according to the number of base features they combine. The simplest feature types—those consisting of values drawn from only one annotation category—may be of direct interest to scholars of classics (e.g. does one author use participles more frequently than another?). All were retained except for the morphological category of degree of comparison in both target word and parent word. As not all adjectives admit of degree of comparison, the AGLDT regularly annotates this category only when the word in question is comparative or superlative. Such words are too rare for the feature to survive in further steps in the selection process. For complex feature types made up of combinations of from two to five base features, the most frequent feature types at each level of complexity were retained. Only those complex feature types were kept which occur in at least 7,000 tokens (20%) throughout the feature selection corpus. The result was a total of 991 feature types: simplex, 18/20 (90%); two-plex, 71/185 (38.3%); three-plex, 194/1,140 (18.4%); four-plex, 331/3,804 (8.7%); and five-plex 377/9,910 (3.8%) possible types.

These 991 frequent feature types must be subjected to yet another round of winnowing. Each feature type may be represented in a text by many possible values. For example, the four-plex type that combines the syntactic relation and part of speech of the target word and the same two features of the parent word has 1,707 distinct values in our sample corpus. Not surprisingly for data on natural
language features, these values are very unevenly distributed. A total of 1,360 of the values occur ten times or fewer, and 747 values occur only once. Nor is this particular type unusual in the number of its values: when all 991 types are associated with their annotations, the result is 456,648 unique type–value pairs. This is an intractable number, and the extreme sparsity of the data makes it unlikely that most type–value pairs would be useful for authorship classification, especially of small texts. Thus, all pairs occurring fewer than 1,000 times in the sample set of 35,000 tokens were dropped. This action reduced the number of total features to 898 (of 456,648).

A few examples of type–value pairs may make it easier to picture data generated by this process of combination and retained as independent variables for the classification of the target corpus:

- self-relation==atr
- parent-morph-pos==verb
- self-relation & parent-morph-person & parent-morph-mood==sbj/thirdperson/indicative

The variables look cumbersome, but their rather discursive forms make them easy to read and to interpret. The name for the feature type is composed of the treebank annotation type (e.g. ‘relation’ and ‘morph-pos’) with the prefix ‘self-’ or ‘parent-’ to indicate whether the information comes from the target word or its syntactic ‘antecedent’. If the feature type is complex, with more than one category of annotation, the name of the type is compounded from the names of its component types, separated by ‘_&_’. The value associated with the type is given after the ‘==’ symbol. When the value is complex, its parts are separated by the ‘/’ character. Thus, in the third example above, information is recorded for the syntactic relation of the target word and the person and mood of its parent: the target word is a subject dependent on a third person indicative verb (‘==sbj/thirdperson/indicative’).

5 Classification Testing

The purpose of the experiment is to observe the degree to which classification performance degrades as the size of the texts in the input decreases. The working hypothesis is that data based on shallow and deep syntactic information will allow more accurate classification than the results reported for measures based on vocabulary richness. Thus, at each stage of the testing, each of the ancient Greek works in each corpus was divided into smaller ‘texts’ and these smaller units were classified according to the author. Text sizes ranged from 2,000 to 50 tokens. Increments of decrease were 100 from 2,000 to 100 tokens, with a final drop to 50 tokens.

For the purposes of dividing the works of the corpora, each was treated as a bag of words or tokens. In other words, each token was treated as independent of all others (except in so much as each token was annotated with morpho-syntactic information about its dependency parent); no further account was taken of the context of an individual token in sentence, paragraph, or any other unit of composition. To create the range of texts, segments of the appropriate size were populated with tokens sampled randomly from the annotated treebank file of a particular work. Thus, segments may contain tokens from many ‘parts’ of the original text without regard for their original order. In addition, because hand-annotated data are precious, the size of the segments is approximate. Each segment contains a number of tokens which balances the requirements that the groups of tokens be of nearly equal size and also as close as possible to the target cardinality. For example, the input file of Oration 1 of the Athenian politician Aeschines contains 15,948 tokens. To generate 1,000-token texts, it is divided into sixteen segments, and the deficit of fifty-two tokens is distributed among the groups, leaving an average segment size of 996.75. These approximations in segment size do not pose a problem for the classification test, since they always produce segments that are smaller than the nominal size and which therefore would be expected to lead to classification results with more errors than would segments of the target size. In any event, the difference between nominal and actual size approaches zero as segment size decreases: the corpus-wide mean actual token number for 2,000-token samples is 1,909.7 (mixed corpus); for 100-token samples it is 99.7.
The next step after the partition of the input files into randomly populated sets of tokens is to use the type–value pairs for each token to calculate relative frequency, in each segment, for all 898 features. The result is a matrix in which each row represents a different text segment and each column gives the relative frequency for a specific type–value pair. The task for the classification algorithm is to generate a model that can associate each row of data with the appropriate author.

To train and test the classification models, the data for each text size were partitioned: 10% of the segments (matrix rows) were reserved for testing, and the remaining 90% was used to develop the model. Inclusion of a segment in the training or testing set was random. However, the amount of text by individual authors in the corpus varies a great deal, and this situation could lead to strange results—e.g. the model might restrict its predictions to 'Iliad author', 'Odyssey author', or 'Polybius' with high accuracy, since these are by far the most common authors. Thus, the random assignment into training and test sets was guided by assigning a selection probability to each segment so that each of the thirteen authors was represented by approximately one-thirteenth of the segments in the test set.

Studies of text classification have found many algorithms valuable for this purpose (Stamatatos, 2009), and most have been packaged for scholars who are not computer experts. In this study, linear classification was chosen because of its ability to handle a large number of observations and variables. More specifically, we used the LiblineaR package for R (Fan et al., 2008; Helleputte et al., 2017). This package offers a range of linear methods. Pretesting on small samples revealed that two methods, L2-regularized logistic regression and L2-regularized L2-loss support vector classification, were most accurate (and quickest).13 Both methods were used for input texts of each size with very little difference in results.

Monte Carlo subsampling (Simon, 2007) served to validate the results of the classification testing. Subsampling was applied at two levels. At each input text size, the populating of the segments with randomly selected tokens was carried out ten times. For each of these generations of text segments, 100 random partitions into a training set and a test set were made. The results of the 1,000 tests at each text size were averaged. These steps minimize the effects of the make-up of particular segments or of their inclusion or exclusion from the training/testing groups.

6 Results for the Mixed Corpus

The results for classification according to the size of the text segments are given in Table 1. Although there are many metrics for judging the success of a classifier, these are not germane to this study, since we are not primarily concerned with the particular kinds of mistakes the tests produce (e.g. which authors are likely to be confused). Rather, the focus of this investigation is the general feasibility of classification based on dependency syntax without the confounder of vocabulary data. Therefore, the simplest measure of accuracy is used here: percentage of successful attributions.

These results are clearly very encouraging. Multiclass categorization on the basis of syntactic features alone seems feasible, even for small texts.

7 Possible Confounders

However, we must not take these results at face value without further examination. The input corpus contains poetry and prose, divided among the genres of epic, drama, history, judicial oratory, biography, and essays of miscellaneous subject. In addition, the texts in the AGLDT are the product of many annotators, a fact that introduces the possibility that the classifier is detecting characteristics of the annotators rather than the authors. Certainly, it is reasonable to assume that generic and annotator differences may contribute in a significant way to the accuracy of the method.

Given the amount of annotated data available when this experiment was designed and set on its course, little could be done to avoid the ramifications caused by using such a diverse corpus. There were too few texts in any given genre to allow for a multiclass classification problem within that genre. On the other hand, an attempt was made to mitigate
the effect of different annotators by judicious design
of the variables to be used for classification. In par-
ticular, it is hoped that including many variables
based on combinations of morphological features
would provide objectivity. It is not unusual to find
annotators disagreeing about the correct depend-
cy relationship for a token, or even about which
token is its parent. It is rare, however, for qualified
annotators to differ over, e.g. the tense of a verb.
Fortunately, the AGLDT continues to expand
and a large body of prose texts has recently been
added. This addition allows us to address directly
the issue of genre and annotators as confounders.

8 The Prose Corpus: Description
and Results

The websites associated with the AGLDT now con-
tain enough prose texts to support a thirteen-class
classification, the same number of classes as in the
mixed corpus presented above. In addition, the
new texts were annotated by the same scholar res-
ponsible for the prose texts included in the mixed
corpus, as described above. Thus, we are able to
reapply our method to a corpus of 328,482 tokens
that is controlled for the prose-poetry distinction as
well as annotator variability. The results are pro-
vided in Table 2.

| Text size | Total ‘Guesses’ | Support vector classifier | Logistic regression classifier |
|-----------|----------------|---------------------------|-------------------------------|
|           |                | Total errors | Accuracy (%) | Total errors | Accuracy (%) |
| 2,000     | 30,000         | 5            | 99.98        | 0            | 100          |
| 1,900     | 32,000         | 7            | 99.97        | 4            | 99.98        |
| 1,800     | 33,000         | 1            | 99.99        | 5            | 99.98        |
| 1,700     | 36,000         | 8            | 99.97        | 10           | 99.97        |
| 1,600     | 37,000         | 16           | 99.95        | 14           | 99.96        |
| 1,500     | 40,000         | 5            | 99.98        | 22           | 99.94        |
| 1,400     | 43,000         | 20           | 99.95        | 16           | 99.96        |
| 1,300     | 46,000         | 16           | 99.96        | 30           | 99.93        |
| 1,200     | 50,000         | 21           | 99.95        | 42           | 99.91        |
| 1,100     | 54,000         | 30           | 99.94        | 28           | 99.94        |
| 1,000     | 59,000         | 52           | 99.91        | 60           | 99.89        |
| 900       | 65,000         | 93           | 99.85        | 124          | 99.80        |
| 800       | 74,000         | 98           | 99.86        | 88           | 99.88        |
| 700       | 84,000         | 130          | 99.84        | 231          | 99.72        |
| 600       | 98,000         | 279          | 99.71        | 259          | 99.73        |
| 500       | 117,000        | 486          | 99.58        | 468          | 99.60        |
| 400       | 147,000        | 808          | 99.45        | 911          | 99.38        |
| 300       | 195,000        | 2,130        | 98.90        | 2215         | 98.86        |
| 200       | 292,000        | 5,881        | 97.98        | 6,960        | 97.61        |
| 100       | 584,000        | 33,281       | 94.30        | 33,085       | 94.33        |
| 50        | 1,166,000      | 166,188      | 85.74        | 164,264      | 85.91        |
mind, it is valuable to examine the results of classifying texts of a single genre, narrative history.

9 The History Corpus: Description and Results

A subset of the prose corpus, the history corpus includes works by seven authors. The corpus is comprised of 200,927 tokens with nonsyntactic punctuation removed. Table 3 provides the results for this monogeneric corpus.

Because of the difference in the number of authors, these results are not directly comparable to those of the larger corpora. It is nonetheless clear that the method examined here is apparently resistant to the most worrisome possible confounders.

10 Discussion

Previously published studies (Eder 2015, 2017) have found that 2,000 words is an approximate lower limit for text size for successful multiclass classification based on measures of vocabulary richness. This limit has been found to apply to texts in ancient Greek as well as several other languages. In this investigation, using only syntactic data from dependency trees, no such lower limit has been discovered. As is to be expected, accuracy does decrease as input text segments become smaller, but the decline is slow and fairly steady. Surprisingly, the slope of the rate of decline apparently flattens for smaller input. This curve is quite different from that reported by Eder (2015), where shrinking text size induces a clear flexion point: after following a gentle slope above c. 3,000 tokens, classification accuracy ‘drops off a cliff’. In contrast, classification through dependency syntax features seems robust even with respect to very small input texts. In fact, an exploratory test of ten-token segments, using the mixed corpus and the same variables described above, has produced an accuracy rate of 50.8% (see below).

The results described here immediately raise the question of why this set of syntactic variables has apparently proven much more effective than features based on vocabulary. To answer this question decisively would require a detailed examination of the effects of each variable and their combinations. Such work is beyond the bounds of this study, which is limited to testing the hypothesis that syntactic data in small text categorization may be more valuable than is generally recognized. At the same time, it is reasonable to suggest that a significant part of the advantage which syntactic features may hold is simply due to the greater richness of syntactic data.

Table 2 Classification of the prose corpus

| Text size | Total ‘Guesses’ | Total errors | Accuracy (%) |
|-----------|----------------|--------------|--------------|
| 2,000     | 17,000         | 0            | 100          |
| 1,900     | 18,000         | 0            | 100          |
| 1,800     | 18,000         | 0            | 100          |
| 1,700     | 20,000         | 2            | 99.98        |
| 1,600     | 21,000         | 0            | 100          |
| 1,500     | 22,000         | 0            | 100          |
| 1,400     | 24,000         | 1            | 99.99        |
| 1,300     | 25,000         | 0            | 100          |
| 1,200     | 28,000         | 1            | 99.99        |
| 1,100     | 30,000         | 1            | 99.99        |
| 1,000     | 33,000         | 0            | 100          |
| 900       | 37,000         | 4            | 99.98        |
| 800       | 41,000         | 0            | 100          |
| 700       | 47,000         | 17           | 99.96        |
| 600       | 55,000         | 38           | 99.93        |
| 500       | 66,000         | 157          | 99.76        |
| 400       | 83,000         | 286          | 99.65        |
| 300       | 110,000        | 662          | 99.39        |
| 200       | 165,000        | 2,705        | 98.36        |
| 100       | 329,000        | 18,466       | 94.38        |
| 50        | 657,000        | 103,748      | 84.20        |

Table 3 Classification of the history corpus

| Logistic regression classifier |
|-------------------------------|
| Text size | Total ‘Guesses’ | Total errors | Accuracy (%) |
|-----------|----------------|--------------|--------------|
| 500       | 40,000         | 4            | 99.99        |
| 400       | 50,000         | 85           | 99.83        |
| 300       | 67,000         | 130          | 99.80        |
| 200       | 100,000        | 738          | 99.26        |
| 100       | 201,000        | 6,071        | 96.97        |
| 50        | 402,000        | 41,801       | 89.60        |
relatively large number of data points per token. There are, of course, many fewer syntactic categories than content word types. As a result, although each syntactic category allows a choice among a number of possible values, many type–value pairs are very common indeed. This is true even when syntactic types are combined. Recall that we limited our set of features to only those type–value pairs which occurred in our sample corpus at a rate equal to or higher than 2.85% (1,000 times in a 35,000-token sample). When the feature set chosen on this basis was populated from the full mixed corpus, the least frequent datum was associated with 2.20% of the total number of tokens. Not surprisingly, this ‘rare’ type–value pair is a combination of five morpho-syntactic categories, the largest size of combinations considered in our variable selection: in the mixed corpus, 12,853 tokens (from a total of 582,487) are dependent on a token which is ‘present indicative active 3rd person singular’. Thus, the least common item in our feature set of 898 variables is dependency on a verb of this very frequent morphology.

If we look, in contrast, at the frequency distribution of vocabulary and select the 900 most common word types in the preserved corpus of ancient Greek, the least frequent words in that selection have a frequency of 0.01%. Thus, these word types are 220 times less common than the corresponding morpho-syntactic feature. At the head of the distribution, the difference is also sharp, if not involving orders of magnitude. There are apparently only three-word types in the entire corpus which occur more frequently than 2%: ο (‘the’, 12.6%), καὶ (‘and’, 5%), and ἄν (‘and’, 3.4%). By way of comparison, the rank and frequency of the top five morpho-syntactic variables are given in Table 4.

Two comments are in order about these data. First, the distribution does not look Zipfian. This fact should not be worrying, since our set of variables does not represent a natural linguistic category, but rather is a conglomeration. The distribution of these variables is therefore the result of a combination of slices from the presumably-Zipfian distributions of other phenomena. Second, the variables used in this study are often collinear. For example, the features ranking fourth and fifth in Table 2 are essentially the same variable, since verbs are the only part of speech in Greek with the category ‘voice’. Collinearity is not a problem for the classification algorithms used in this investigation. Thus, such variables were not culled from the feature set as part of the selection process. On the other hand, collinear data will need to be accounted for when the influence of individual variables is assessed.

Be that as it may, it is plausible to assume that the relatively high frequency of these morpho-syntactic features is a principal cause for the viability of classification on this basis. In particular, these features help us to avoid the pitfalls of data sparsity. The average number of features per token in the mixed corpus is 48.7 (with a maximum of 212). When tokens associated with so many variables are assembled into the randomly selected bag-of-words ‘texts’, this process results in fairly dense data: one example matrix representing 100-token samples from the mixed corpus has 5,840 rows and 898 columns for 5,244,320 cells. Of these cells, 281,539 have a value of zero. Thus, this matrix has a sparsity ratio of only 5.37%. Under these circumstances, the classification accuracy of 94% is perhaps not surprising. If we examine sparsity data generated from the prose corpus, we can see that the rate of change in sparsity between segment sizes is fairly similar to that observed for classification accuracy (Table 5).

For both sparsity and classification error, the data curve remains quite flat until roughly the 600-token segment mark for the error rate and the 300-token mark for sparsity. It is noteworthy that between the 100- and the 50-token segments the sparsity rate increases by a factor of 2.817 and

| Type–value pair | Rank | Frequency (%) |
|-----------------|------|---------------|
| parent-morph-pos==verb | 1 | 37.7 |
| self-morph-number==singular | 2 | 37.6 |
| parent-morph-number==singular | 3 | 37.3 |
| parent-morph-voice==active | 4 | 26.6 |
| parent-morph-pos & parent-morph-voice==verb/active | 5 | 26.6 |

Table 4 The most common syntactic variables
the error rate by a factor of 2.813. More detailed study of such a correlation may bear fruit in possible future attempts to identify the reasons for the success of our approach. Whatever the precise cause, the features in the data preserve an unexpected amount of information, even when the input ‘texts’ are much smaller than are likely to be of interest to scholars working on real-world problems. For example, we performed exploratory experiments to see if the morpho-syntactic features would continue to be useful with input segments of ten tokens or fewer. Classification was still quite successful:

- Ten-token segments: 50.8% accuracy
- Five-token segments: 38.8% accuracy
- Two-token segments: 23.1% accuracy.

The matrix for the two-token texts from the mixed corpus has a sparsity ratio of 89.62% and, with a mean ‘No Information Rate’ of 8.16%, produces an accuracy that approaches three times that of random chance.

These accuracy rates with ‘micro-texts’ may be surprisingly good when compared with chance, but they will be unlikely to satisfy scholars trying to solve real classification problems. However, it is worth remembering that these results are from multiclass categorizations that have been chosen for this study as a ‘harder’ problem than binary classification. Yet, in the real world of authorship classification, many questions, even if they involve the ultimate goal of identifying the most likely author from a large group of possibilities, can be usefully recast as binary problems: is a given writer likely to be the author of a particular text?

If the problem can be put in those terms, the morpho-syntactic approach outlined here may prove satisfactory with quite small texts. We may very briefly examine this possibility, first by looking at a toy example, then at a more realistic one.

Our mixed corpus includes both prose and verse texts, with proportions being roughly equal. We may therefore see how well this distinction can be recognized with very small input texts. Using the logistic regression algorithm on the morpho-syntactic input results in the following accuracies of distinguishing poetry from prose:

- Ten-token segments: 97.1%
- Five-token segments: 91.6%
- Two-token segment: 81.1%.

A more interesting test involves the ability to reliably say ‘yes’ or ‘no’ to the question, ‘Did author x write this text?’ The historians Herodotus and Thucydides may serve as the examples drawn from our corpus. The categories will thus be ‘Herodotus’ and ‘other’ and ‘Thucydides’ and ‘other’, respectively. In this test of the mixed corpus, the input texts will contain twenty tokens. This size approximates the average sentence length in the corpus, and texts shorter than a single sentence often cannot be annotated for dependency syntax. The results using logistic regression are good:

- Twenty-token segments, Herodotus versus other: 80.9%
- Twenty-token segments, Thucydides versus other: 86%.

It will not surprise anyone familiar with the syntax of the Greek historians that Thucydides is somewhat easier to distinguish than Herodotus. Thucydides is

| Text size | Mean sparsitya (%) | Classification errors (%) |
|-----------|-------------------|---------------------------|
| 2,000     | 1.71              | 0                         |
| 1,900     | 1.72              | 0                         |
| 1,800     | 1.72              | 0                         |
| 1,700     | 1.73              | 0.02                      |
| 1,600     | 1.73              | 0                         |
| 1,500     | 1.74              | 0                         |
| 1,400     | 1.74              | 0.01                      |
| 1,300     | 1.75              | 0                         |
| 1,200     | 1.76              | 0.01                      |
| 1,100     | 1.77              | 0.01                      |
| 1,000     | 1.78              | 0                         |
| 900       | 1.79              | 0.02                      |
| 800       | 1.81              | 0                         |
| 700       | 1.83              | 0.04                      |
| 600       | 1.86              | 0.07                      |
| 500       | 1.90              | 0.24                      |
| 400       | 1.96              | 0.35                      |
| 300       | 2.07              | 0.61                      |
| 200       | 2.42              | 1.64                      |
| 100       | 4.72              | 5.62                      |
| 50        | 13.32             | 15.80                     |

aThe average for the ten iterations of token partitioning for each segment size.
well-known as a syntactic outlier in many respects. In any case, a successful classification rate of about eight-tenths may persuade scholars in the field to give some attention to computational evidence.

We may assume that the results presented in this study represent a lower limit for the expected effectiveness of this approach. The variables used to achieve the recorded accuracies were selected in the most naïve way: first, frequency of variable type and, second, frequency of type–value pairs. It is reasonable to expect that a more systematic examination of the variables will identify the most and least useful features in the present set. This knowledge may improve results through the elimination of noise and by suggesting additional type–value pairs to be included.

More important perhaps may be the incorporation of other strands of morpho-syntactic information. At a very basic level, language may be thought of as the product of (at least) two ‘orders’: a hierarchical order represented in a syntax tree, and a linear order given by the chronological sequence of words in a sentence. A weakness of dependency treebanks such as those used in this study is that they take no account of word order. We might therefore expect that the introduction of data about word order into our features set would lead to improvement. This expectation seems to be borne out in provisional testing. A straightforward way to include some word order information is to use bigram data alongside the variable set explained above. Morphosyntactic bigram type–value pairs contain data about the target word and/or the following word. Examples may make this structure clear:

- bigram-morph-case_&_self-relation==genitive/atr
- bigram-morph-case_&_bigram-parent-morph-pos==nominative/verb.

The first variable indicates that the word following the target word is genitive, while the target word itself is syntactically the attribute of its parent word. The second variable indicates that the case of the word after the target word is nominative and it (the following word) is dependent on a verb. When the most frequent of these bigram features are added to the original data set for the mixed corpus, accuracy for fifty-token segments increases from 85.9% to 90.9%. Thus, we may be cautiously optimistic that experimentation with different kinds of morpho-syntactic variables may bring significant improvement.

Categorization of texts of all sizes on the basis of combinations of deep and shallow linguistic features looks quite promising, but many details remain to be worked out. The experiments in this investigation were based on random bag-of-words input, a method quite frequent in the literature. Scholars are more likely to be interested in classifying continuous texts, so the value of morpho-syntactic variables must be tested with this kind of input. Another question to be considered is the amount of training data available. All authors in this study were represented by at least c. 20,000 tokens of text, but in real-world classification problems, small target texts often come with limited training material for the authors of interest. The extent to which the method presented here is sensitive to the effects of training data size must be established. Positive assessment of a morpho-syntactic approach to classification of small texts must recognize its most serious drawback: the necessary syntactic annotation is expensive and time-consuming. The set of available texts is not large, and those wishing to apply this method of attribution to other material will need to create their own corpus. Fortunately, automated dependency parsing is becoming more feasible and may soon help to overcome this obstacle.

In addition, while the method presented here gives good results when applied to a morphologically rich language such as ancient Greek, it remains to ascertain whether such morpho-syntactic data will also prove valuable for classification of short texts in less morphologically complex languages such as English.

In sum, in spite of certain limitations, morphosyntactic information can form a strong basis for classification problems involving an unusual range of text sizes. While this investigation looked at classification by morpho-syntactic variables alone, it is reasonable to suppose that, as data of this kind becomes more readily available, they can be usefully incorporated into many types of analyses to benefit researchers working on a wide variety of questions in natural language processing.
References

Bamman, D. and Crane, G. R. (2011). The ancient Greek and Latin dependency treebanks. In Sporleder, C., van den Bosch, A., and Zervanou, K. (eds), Language Technology for Cultural Heritage. Berlin: Springer, pp. 79–98.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. Digital Scholarship in the Humanities, Vol. 0, No. 0, 2019.

Eder, M. (2017). Short samples in authorship attribution: a new approach. Digital Humanities 2017: Conference Abstracts. Montreal: McGill University, pp. 221–24. https://dh2017.adho.org/abstracts/341/341.pdf (accessed 22 May 2019).

Fan, R.–E., Chang, K.–W., Hsieh, C.–J., Wang, X.–R., and Lin, C.–J. (2008). LliblineaR: Linear Predictive Models Based on the Lliblinear C/C++ Library. R package version 2.10–8. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004), pp. 611–17. https://dl.acm.org/citation.cfm?doid=1220355.1220443 (accessed 22 May 2019).

Gorman, R. J. and Gorman, V. B. (2016). Approaching questions of text reuse in ancient Greek using computational syntactic stylometry. Open Linguistics, 2: 500–10.

Helleputte, T., Gramme, P., and Paul, J. (2017). Liblinear C/C++ Library. R package version 2.10–8. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. Digital Scholarship in the Humanities, 22(4): 405–17.

Koppel, M., Schler, J., Argamon, S., and Winter, Y. (2012). The ‘Fundamental Problem’ of authorship attribution. English Studies, 93(3): 284–91.

Kuhn, M. (2016). Caret: Classification and Regression Training. R package version 6.0–73. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Luyckx, K. and Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 513–20. https://dl.acm.org/citation.cfm?id=1599146 (accessed 22 May 2019).

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing, 26(1): 35–55.

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona–Hernández, L. (2012). Syntactic dependency-based N-grams as classification features. Lecture Notes on Computer Science, 7630: 1–11.

Simon, R. (2007). Resampling strategies for model assessment and selection. In Dubitzky, W., Granzow, M., and Berrar, D. (eds.), Fundamentals of Data Mining in Genomics and Proteomics. Boston: Springer, pp. 173–86.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3): 538–56.

Stamatatos, E. (2016). Authorship verification: a review of recent advances. Research in Computing Science, 123: 9–25.

Stamatatos, E. (2017). Masking topic-related information to enhance authorship attribution. Journal of the Association for Information Science and Technology, 69(3): 461–73.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Proceedings of the Second Asia Conference on Asia Information Retrieval Technology (AIRS '05). Berlin: Springer. pp. 174–89.

Notes

1 http://perseusdl.github.io/treebank_data/ (accessed 22 May 2019) and https://perseids-publications.github.io/gorman-trees/ (accessed 18 August 2019). Note that some files are common between the two sites. The main Perseids Project homepage is at https://www.perseids.org/.

2 Aeschines, Oration 1; Aeschylus, Agamemnon, Eumenides, Libation Bearers, Persians, Prometheus Bound, Seven Against Thebes, Supplicants; Demosthenes, On the Crown, Philippic 1; Diodorus Siculus, Historical Library Book 1; Herodotus, Histories Book 1; Homer, Iliad, Odyssey; Lysias, Orations 1, 12, 13, 14, 15, 19, 23; Plutarch, Life of Alcibiades, Life of Lycurgus; Polybius, Histories Book 1-2; Sophocles, Ajax, Antigone, Electra, Oedipus the King, Trachinia Thucydides, History of the Peloponnesian War Book 1, Book 3, 31-40; Xenophon, Cyropaedia Book 1, Hellenica Book 1. Note that the Iliad and the Odyssey are generally considered to be by
different authors, in spite of the traditional attribution to a single ‘Homer’. Thus, in this study a separate class label has been assigned to each.

1 I would like to thank an Anonymous Reviewer for emphasizing the seriousness of this concern.

2 In general, individual texts were rejected from the target corpus only on the basis of size. If the texts attributable to an author did not include approximately 20,000 tokens, they were set aside. For example, the Hesiodic Theogony, Works and Days, and Shield were excluded since it is safest to consider them by different authors and the resulting text size for each author is too small. This size limit was based on the need to allow for a 90–10% partition between training and testing sets with 2,000-token segments. The work of Athenaeus, although of sufficient length, was excluded because it is essentially a patchwork of passages drawn from other authors. In a similar vein, only Polybius Books 1 and 2 were included; the other annotated texts of that historian come from books transmitted indirectly and therefore subject to abridgement and paraphrase. On the other hand, directly transmitted texts of dubious authorship were not excluded. It was assumed that grouping possibly ‘misattributed’ works with the genuine works of an author would lead the algorithm to have more difficulty modelling the syntactic signature of that author. Thus, the mixing of genuine and doubtful texts might be expected to decrease the accuracy of the classification. Such an effect would not cast doubt on the usefulness of the method presented here.

3 On the distinction, see Gamon (2004).

4 Strictly speaking, the inclusion of the dependency relation of the parent word introduces information about a third generation of words, since this datum is not a feature of the parent word per se, but a description of the relationship between that word and its parent. Some further information such as part-of-speech can be implied about the third generation by the relation data (an SBJ parent entails a verbal grandparent), but no morphological information for the grandparent is explicitly included. My thanks to an anonymous reviewer for pointing out these implications. It is worth noting that removal of all variables including the dependency relation of the parent does not seem to materially affect the result of the classification. An exploratory test of 200-token segments produced an accuracy of 0.9926 with the third-generation information and 0.9925 without.

5 Often the data indicate ‘not applicable’, since some categories are limited to verbal forms, others to nominal forms, and so on.

6 These variables are essentially an elaboration of the syntactic ‘sn-grams’ (syntactic n-grams) discussed in Sidorov et al. (2012).

7 In this step, a number of texts were included which were later dropped from the classification test itself as unsuitable. A total of 35 texts were sampled for a feature selection corpus of 35,000 tokens.

8 The quirks of the AGLDT annotation are reflected in some type–value pairs. The AGLDT schema allows use of punctuation marks as parents of coordinated or apposed elements. Thus, variables including punctuation were admitted to the feature set for classification of the mixed corpus of poetry and prose. All tokens in the AGLDT are assigned a dependency relation with a parent. This is true even for, e.g. the main verb of a sentence, which is assigned a fictitious dependency relation tagged as ‘root’. Type–value pairs with the value ‘root’ were also admitted to the feature sets for all corpora. In contrast, this root node has no morphology and therefore never appears in any data category except dependency relation. More generally, a token must have ‘positive’ values in all components of a combined variable in order to be included in the count of that variable. For example, since infinitives as a rule have no grammatical person, the pre-processing algorithm assigns ‘NA’ (‘not applicable’) to the associated value for each infinitive token. When relative frequencies for variables including ‘person’ are calculated, all infinitives are then dropped. A somewhat complicated instance is provided by the handling of ellipses in the AGLDT. An annotator may add an unlimited number of tokens to a sentence to allow a full graphing of its syntactic structure. Usually, ellipses are annotated with no morphological information, except occasionally part-of-speech. Thus, their putative morphology does not affect this study. On the other hand, a dependency relationship is regularly annotated between an ellipsis and its parent and descendants. These relationships are usually no more difficult to identify than those not involving ellipses, and for this reason information about them is included in our variables. Likewise, the morphological information from words sharing an edge with ellipsis nodes is admitted as a matter of course, since annotating morphology is rarely hindered by connection with an ellipsis.

9 This comes to 991 of 14,973 total feature types (6.6%). There are many fewer total types than we would expect based on the combination of one to five elements (21,699). The rules of Greek grammar disallow many such combinations.

10 The segments were created using the ‘createDataPartition’ functions of the caret package (Kuhn 2016) of the R programming language (R Core Team 2018).
13 Regularization is a method to prevent the overfitting of a model to the training data. Here, it reduces the influence of any particular independent variable. A loss function measures the ‘wrongness’ of a prediction. A machine learning algorithm seeks to minimize this measure to optimize performance. In an L2 function, the relevant quantity is squared. The details (Fan et al., 2008) are not important for the purpose of this study.

14 http://perseusdl.github.io/treebank_data/ (accessed 22 May 2019) and https://perseids-publications.github.io/gorman-trees/ (accessed 18 August 2019). The added texts meeting our minimum size requirement are Antiphon, Orations 1–2, 5–6; Appian, Civil Wars Book 1; Dionysius of Halicarnassus, Roman Antiquities Book 1; Flavius Josephus, Jewish War Book 7; Plutarch, On the Fortune of the Romans, On the Fortune or Virtue of Alexander the Great. Lysias 15 was dropped as possibly spurious.

15 Support Vector classification was not performed for this corpus, since for the mixed corpus results were not interestingly different for the two algorithms.

16 The prose corpus is slightly ‘cleaner’ than the mixed corpus. All nonsyntactic punctuation has been removed from the newer corpus: any column with a variable containing such a value was dropped. In addition, any row representing such a punctuation token was dropped before calculation of relative frequencies. In AGLDT annotation, syntactic punctuation marks are given the relation tag ‘COORD’ or ‘APOS’. These tokens were retained.

17 Recall that the corpus is randomly partitioned ten times into segments of each size, and then the segments of a particular partitioning are classified 100 times, each time with a different random selection of training and test segments. Occasionally, the partitioning process generates segments that are difficult for the model to handle correctly. Such segments may be repeatedly misclassified during the 100 iterations. For example, in the 600-token classification, a particular segment of Josephus appeared in the test set nine times. It was classified correctly twice and seven times was misidentified as a Plutarch passage. These seven mistakes can be counted as one ‘error type’.

18 Josephus as Plutarch and Xenophon as Lysias. The intra-genre error types are Demosthenes as Aeschines, Aeschines as Demosthenes (two different partitions), Lysias as Antiphon, Xenophon (Hellenica) as Thucydidides, Antiphon as Lysias, Demosthenes as Antiphon, and Diodorus as Polybius.

19 Appian, Diodorus, Dionysius, Herodotus, Josephus, Polybius, and Thucydidides. Xenophon’s Hellenica was excluded since the available annotated text is too short.

20 Since the tests of the other corpora show practically no errors with segments larger than 500 tokens, classification of larger segments was not undertaken.

21 Eder (2015) examines evidence for English, German, Ancient Greek, Hungarian, Latin, and Polish.

22 For example, in the mixed corpus, the error rate at 2,000-token segments is 0.02%, while for 500-token segments the figure is 0.42%. Thus, the error rate has increased by a factor of 21. On the other hand, 200-token segments show 2.02% error, and 50-token segments 14.26%. Here the increase is only a factor of 7, although the ratio between the size of the input segments is the same. Of course, some of the difference may be explained by the relatively high variance in tests of larger segments due to the small size of the training/test sets.

23 The figures come from the word frequency tool in the Perseus under PhiloLogic website (http://perseus.uchicago.edu/, accessed 5 January 2019). These frequencies are calculated from something over 5 million words distributed among 175 texts. Faute de mieux, we may accept the statistics for word distribution as a representative sample for all ancient Greek literature.

24 Koppel et al. (2012) discuss the relationship between single-label multiclass classification problems and the more basic ‘binary authorship verification problem: determine if the given document was written by a specific author or not’ (p. 286, emphasis in original). The authors consider this binary reduction to be the ‘fundamental problem’ of authorship classification and argue that solving the basic binary problem will lead to the solution of ‘any of the standard authorship attribution problems’ (p. 284).

25 To avoid the ill effects of data in which the distribution between the target author and ‘other’ is very unbalanced, an appropriate number of 20-token segments were randomly selected from among all nontarget segments, so that the number of segments representing the two classes would be approximately equal after the random partition into training and testing sets. In addition, all nontarget authors were represented by an approximately equal number of segments.

26 The Support Vector algorithm was used. The 1,283 type–value pairs which occurred at least 1,000 times in the 35,000-token sample corpus were selected. The resulting full corpus data had a mean of 113.8 type–value pairs per token. For randomly generated 20-token segments, the data matrix had a sparsity ratio of 42.2%.

27 For a study based on function words in English documents, see Zhao and Zobel (2005).