Computers and the Humanities **36:** 411–430, 2002.

Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works

J.A. SMITH[1] and C. KELLY[2*]
[1]*Department of Classics & Humanities, San Diego State University, San Diego, CA 92120, USA*
E-mail: jasmith@mail.sdsu.edu
[2]*Department of Mathematical and Computer Sciences, San Diego State University, San Diego, CA 92182, USA*
E-mail: kelly@math.sdsu.edu
*(*author for correspondence)*

Abstract. The measure of the lexical richness of literary texts as a tool in the comparative analysis of literary style has been hampered by the problem of the inequality of text lengths within and between literary corpora. This paper proposes an empirical method of description of lexical richness by averaging measures on multiple chunks of text of a standard length within a literary work or corpus. A work's average vocabulary richness, average portion of *hapax legomena* of the corpus from which it derives, and average repetition of frequently appearing vocabulary may then characterize that work relative to other works partitioned along with it. This method reveals the possibility of significant variance of these measures of vocabulary among works of a single author's corpus and warns against the notion of some absolute authorial stylistic character. We apply this method of vocabulary averaging to the corpora of three playwrights from classical antiquity whose works are chronologically rankable: Euripides, Aristophanes, and Terence. We look for trends in vocabulary richness over time, which we posit functions as an indicator of progressively changing authorial ability or inclination. This method then holds the potential of predicting dates for undateable or tenuously dated works within a corpus of otherwise securely dated texts. From the results derived, a relatively late date for the composition of the redrafted version of Aristophanes' *Clouds* appears likely; we predict an early composition date for the redraft of Terence's *Hecyra* (and thus are inclined to think that the playwright did very little redrafting); and finally we find Euripides' *Electra* and *Supplices* exhibiting vocabulary characteristics of extremely late composition and we predict dates much later than those assigned based on metrical considerations.

Key words: chronology, hapax, prediction, vocabulary, Yule's K

1. Introduction

Measures of the lexical richness in literary texts have long enticed stylometricians with the promise of providing some comprehensive index of authorial style. The hope is that a quantification of either a text's complete vocabulary or the number of rarely occurring word types can numerically express the linguistic complexity

or simplicity of one text relative to other texts and give an indication of the lexical similarity or difference between authors or between sections of a single author's literary output. Yet the numerical characterization of lexical richness is dependent on text length and a formidable obstacle to acquiring such an index of style arises in the natural inequality in the length of works of literature, to say nothing of the differing sizes of entire literary corpora. One approach to overcoming this obstacle has been to invent statistical methods to express the vocabulary richness of a given text that are text-length independent (Tweedie and Baayen, 1998).

We propose the following empirical approach to surmount the problem of text-length dependence of lexical measures. We choose a standard unit of text length on which to calculate our measure of lexical richness. This standard unit of text length is chosen to be sufficiently small so that a large block of literary text can be split into many non-overlapping chunks of the standard length. The average of the measures calculated on a block then gives a new measure of lexical richness that is not dependent on the text length. For an author whose complete corpus of writing consists in many individually published pieces, the stylometrician, partitioning text into equal chunks across the entire corpus and grouping by individual works, can thus describe not only the average vocabulary richness of the corpus but of the individual works within the corpus.

In this paper, we pursue this approach to expressing vocabulary richness in an effort to evaluate the extent to which the stylistic characteristic of vocabulary richness holds constant across a writer's corpus. We have discovered substantial variance in the average vocabulary richness across sections of some of the literary corpora we examined, which has implications for methods that determine authorship based on such measures. Additionally, we have found evidence suggesting that change in lexical richness may be correlated to the chronological sequence of publications within a corpus. The very process of composition may affect an author's compositional skills, abilities, and inclinations in such a way as to register as progressive change in vocabulary richness from work to work. If an author's writing style steadily changes over the course of a career, we can model this change and use it to predict the date of composition of an unknown work within a collection of works of known dates.

2. Selection of Texts for Analysis

The works of just eight dramatists have been transmitted to us from classical antiquity in sufficient abundance and in a sufficient state of preservation that we can meaningfully speak of them as the literary corpora of their authors.[1] These corpora afford unique opportunities for observing vocabulary richness perfectly suited to the purposes of this study thanks to the special circumstances of the composition of classical plays. Unlike other literary works, classical plays were composed in (relatively small) discrete units of text, in accordance with highly traditional conventions of genre, conforming to regular restrictions of theatrical realization,

and, once performed and promulgated, not usually liable to their author's redrafting and re-staging.[2] Despite their susceptibility in ancient times to occasional actor interpolation (in later revivals) and to the more pervasive normalizing emendations of editors from the classical, medieval, and modern ages, classical plays can be subjected to analysis under the hypothesis that each play represents a discrete historic moment of literary output in a specific literary genre. Classical playwrights, characteristically prolific in their output,[3] composed a work, saw it produced in a specific generic category of presentation, and moved on to the next work for the next opportunity for production.

Three of these dramatic corpora – the tragedies of Euripides, the "old" and "middle" comedies of Aristophanes, and the "new" comedies of Terence – are of special interest because the plays within these collections can be ranked chronologically, either entirely or to a large extent, by evidence external to the literary texts.[4] Thus it is possible to quantify elements of literary style in the constituent plays of these authors' corpora and determine whether these stylistic elements change significantly over time. In perhaps the most famous case, progressive change in Euripides' habits of metrical resolution have been used to rank his plays of uncertain performance-date against those that can be securely dated (Devine and Stephens, 1981). Oftentimes, the stylometric techniques which derive data in accord with chronological arrangement for one corpus are then applied to those corpora for which meager external criteria for chronological arrangement exist (i.e. those of Sophocles, Plautus, and Seneca), and the results taken as evidence of consistent development through time (Fitch, 1981; Duckworth, 1952). For the purposes of our study, Euripides, Aristophanes and Terence afford us the opportunity to assess whether the lexical richness of their plays is significantly correlated with time.

3. Measures of Lexical Richness

We consider two measures of literary style that appear to be text-length independent (Tweedie and Baayen, 1998) and thus may be calculated on whole plays without using the partitioning method: Yule's constant, K (Yule, 1944) and Zipf's parameter, Z (Orlov, 1983). Additionally, we consider the length of the play (i.e. number of total words) spoken in iambic meters, N.

Yule's constant characterizes the distribution of word usage into a single parameter, the square of the coefficient of variation. In particular, Yule assumes that different words are used in a work at different rates; his constant is calculated as the ratio of the variance of the word usage rates over the mean rate squared. This constant can be calculated by counting the number of words that are used once, twice, thrice, etc., in a text of length N words. We let $V(i, N)$ denote the number of words that are used i times in a text of length N words. The set $\{V(i, N), i = 1, 2, \ldots,\}$ then gives a distribution of word usage with mean proportional to $\sum i\, V(i, N)$ (which is equal to N), and second moment proportional to $\sum i^2\, V(i, N)$. Yule's

constant is calculated as ten-thousand times the ratio of the second moment minus the mean, to the mean squared:

$$K = 10000 \times (\sum i^2 V(i, N)/N^2 - 1/N). \quad (1)$$

Comparing Yule's constant on works by the same author with the same third and forth moments in the word distribution, Yule (1944) remarks that a high value of K means "the author's vocabulary will in use be highly concentrated, concentrated on to those words that are used over and over again." Conversely, a small value of K means "the author's vocabulary will not be so greatly concentrated on to a few special words" (p. 78).

Zipf's parameter, Z, normalizes the vocabulary size, V, to a text-length independent measure using a complicated function of V, N, and p, the frequency of the most common word:

$$V = \frac{Z \times N \times \log(N/Z)}{(N - Z)\log(p \times Z)}. \quad (2)$$

This parameter specifies the text length at which Zipf's law in its simplest form holds.[5] Tweedie and Baayen (1998) interpret Z as a measure of lexical richness; they claim that an increase in Z leads to an increase in V.

For purposes of this study, we define "word" as a unique combination of letters rendering a distinct (though, perhaps, semantically ambiguous) word-form. Our counts represent the results of an unlemmatized collation of words as found in the somewhat raw word-banks of each author's select corpus which we have edited using stream-editing and script techniques to enhance word separation and regularity of orthography.[6] Given that our results are purely of interest in the relative ranking of characteristics within a corpus and not expressions of some absolute value of style, our primary concern in preparation and treatment of the texts was consistency in word division. By "*hapax legomena*" we mean word types with only one token with respect to the corpus under examination. And so our descriptions of the frequency of *hapaxes* in, e.g. Euripides, pertain only to words written in the iambic sections of fourteen of the nineteen plays which bear his name. A collation of all words in all the plays of Euripides would generate a different data set of "*hapaxes*".

Our decision to limit the word counts of each corpus only to words composed in iambic meters, primarily in the spoken dialogue of iambic trimeter, is motivated by our desire to observe our authors composing according to a single set of rules and conventions. We operate under the hypothesis, then, that the changes we are observing in the measures of lexical richness have to do with changing authorial abilities and inclinations in style of composition and not with larger trends of change in the theater (such as the diminishing role of the chorus and choral music through the careers of Euripides and Aristophanes). The significant evolution in metrical rules of the iambics of Euripides, famously used to rank his corpus chronologically, appears to be a by-product of his changing compositional inclinations and not vice versa (Devine and Stephens, 1981).

We consider three different stylistic characteristics when using our partitioning method: the average lexical richness of the partitioned chunks (i.e. the average number of unique words per chunk of N words), the average number of *hapax legomena* per chunk of N words (i.e. word types with only one token in the entire corpus considered), and the average Yule's constant per chunk of N words (Yule, 1944). Suppose a play is broken into n chunks of size N words; in our analyses, we use $N = 300$.[7] Let v_i be the number of unique word types in chunk i, h_i be the number of *hapaxes* in chunk i, and k_i be Yule's constant calculated on chunk i, where $i = 1, \ldots, n$. For each play, we calculate the mean and standard deviation of each measure:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad s_v = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})^2}$$

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i \qquad s_h = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(h_i - \bar{h})^2}$$

$$\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i \qquad s_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(k_i - \bar{k})^2}$$

While the average number of unique words, $\bar{v}$, measures the general diversity of the vocabulary, the average number of *hapax legomena*, $\bar{h}$, more so than the average number of unique words, gives insight into an author's compositional productivity;[8] both $\bar{v}$ and $\bar{h}$ are measures sensitive to activity at the low frequency end of a text's vocabulary. For measuring activity and variability in an author's use of high frequency vocabulary (most typically, the most common function words of the language), we track the average Yule's constant, $\bar{k}$, for the standard chunk size. This average Yule's constant should be similar to Yule's constant calculated on the entire play, K, if this measure is text-length independent, as claimed. One necessary assumption for this measure to be relatively constant over different chunks is that the writing style is constant, so that the use of words can be modeled as a homogeneous random process. Thus, with these three measures we attempt to track the consistency of style at both ends of each author's lexical profile, and overall.

4. Statistical Methods

4.1. Variability of measures across an author's corpora

Using the partitioning method, we can determine whether an author's style remains fairly constant across the plays or whether it varies. An ANOVA analysis tests whether the lexical averages are the same across the plays. In some cases, the normality assumptions of the ANOVA test were not met in the data, and

a non-parametric alternative, the Kruskal-Wallis Rank Sum test was conducted instead.

4.2. Testing for trends over time

The relationship between the lexical averages for each play and the presentation dates of the plays are assessed using linear regression; because of the differing lengths of the plays and hence the differing standard errors of the measures across the plays, weighted least squares was used to estimate the regression equations (see, for example, Draper and Smith, 1981, pp. 108–115) with weights equal to one divided by the estimated variance of the mean. A unique linear regression equation is estimated for each measure:

$$\bar{v} = \hat{\alpha}_v + \hat{\beta}_v t$$
$$\bar{h} = \hat{\alpha}_h + \hat{\beta}_h t$$
$$\bar{k} = \hat{\alpha}_k + \hat{\beta}_k t$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated intercept and slope of the each regression line respectively, and t is the year in which the play was performed. Wald's tests were used to determine whether there was a significant linear change in a measure over time:

$$W = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \tag{3}$$

where $s.e.(\hat{\beta})$ is the estimated standard error of the estimated slope. Confidence bands for the estimated regression line were estimated as in Draper and Smith (1981).

4.3. Prediction of the date of an unknown play

If literary style is consistently changing over time, as determined by the estimated regression equation, then this equation can be used to predict the date a new play of unknown performance date. Since we would like to predict the date of performance from a given measure, this is a problem in inverse prediction. Using the plot of the measure against time and the estimated regression line, we can predict the date of performance of a play with measure Y (this may be $K, Z, \bar{k}, \bar{h}$, or $\bar{v}$ – whichever is consistently changing over time) by drawing a horizontal line across the plot corresponding to the value of Y and determining where this line intersects the estimated regression line. A 95% prediction interval for the estimated date of performance can be determined in a similar way if prediction bands, rather than confidence bands, are also drawn on the plot. The upper and lower limits of the prediction interval are determined from the points of intersection of the horizontal line with the upper and lower prediction bands respectively. Equations for the

Table I. Measures of lexical richness for eleven plays of Aristophanes

Play	**Date**	$\bar{v} \pm s.e.(v)$	$\bar{h} \pm s.e.(h)$	$\bar{k} \pm s.e.(k)$	K	Z	N
Acharnians	425	218 ± 2.9	63 ± 2.5	38 ± 2.3	27	54	4994
Knights	424	215 ± 3.5	52 ± 3.7	46 ± 3.2	36	34	4412
Clouds	($\geq$ 423)	215 ± 2.7	48 ± 3.3	41 ± 2.0	36	48	4856
Wasps	422	225 ± 3.0	57 ± 2.3	33 ± 2.1	29	53	4698
Peace	421	223 ± 2.2	53 ± 2.3	34 ± 1.8	29	49	4406
Birds	414	218 ± 1.4	53 ± 2.1	36 ± 1.3	30	50	5041
Lysistrata	411	215 ± 3.5	54 ± 3.1	37 ± 1.8	30	61	4561
Thesmo.	411	220 ± 2.2	56 ± 3.8	38 ± 2.9	30	50	4440
Frogs	405	218 ± 2.2	51 ± 2.2	39 ± 1.6	33	40	5506
Ecclesiazusae	392	217 ± 1.9	52 ± 1.9	40 ± 2.0	34	46	5651
Plutus	388	216 ± 2.2	46 ± 2.0	43 ± 3.5	37	42	6561

prediction and prediction interval can be written, but these are complicated. We refer the reader to Draper and Smith (1981).

5. Results

5.1. ARISTOPHANES

We analyzed eleven plays written by Aristophanes for which presentation dates are known with a good degree of accuracy. Table I displays these plays, their performance dates, and the measures of interest.

First, ANOVAs were conducted to determine whether the plays differed substantially from each other in terms of the average lexical measures. Both the average number of *hapaxes*, $\bar{h}$, and the average Yule's constant, $\bar{k}$, were found to differ among the plays (p-values = 0.0005 and 0.016, respectively). The average vocabulary, $\bar{v}$, did not significantly differ across the plays (p-value = 0.105).

Next, weighted linear regression equations were estimated to determine whether the average lexical richness increased or decreased significantly over time. The version of *Clouds* as we have it has been redrafted (Meineck, 2000; Rosen, 1997) and thus may display the writing style of his later career. All regression lines were estimated excluding data on *Clouds*, since the date of this play was uncertain. Using this method, the average number of *hapaxes*, $\bar{h}$, was found to decrease over time (see Figure 1). The coefficients of the regression line are displayed in Table IV; there we see that the average number of *hapaxes* is decreasing significantly (p-value = 0.008) over time at a rate of about one per 300 words every four years. Since *Clouds* was rewritten and we should expect it to display attributes of Aristophanes' later works, we calculate a one-sided (rather than two-sided prediction interval) for

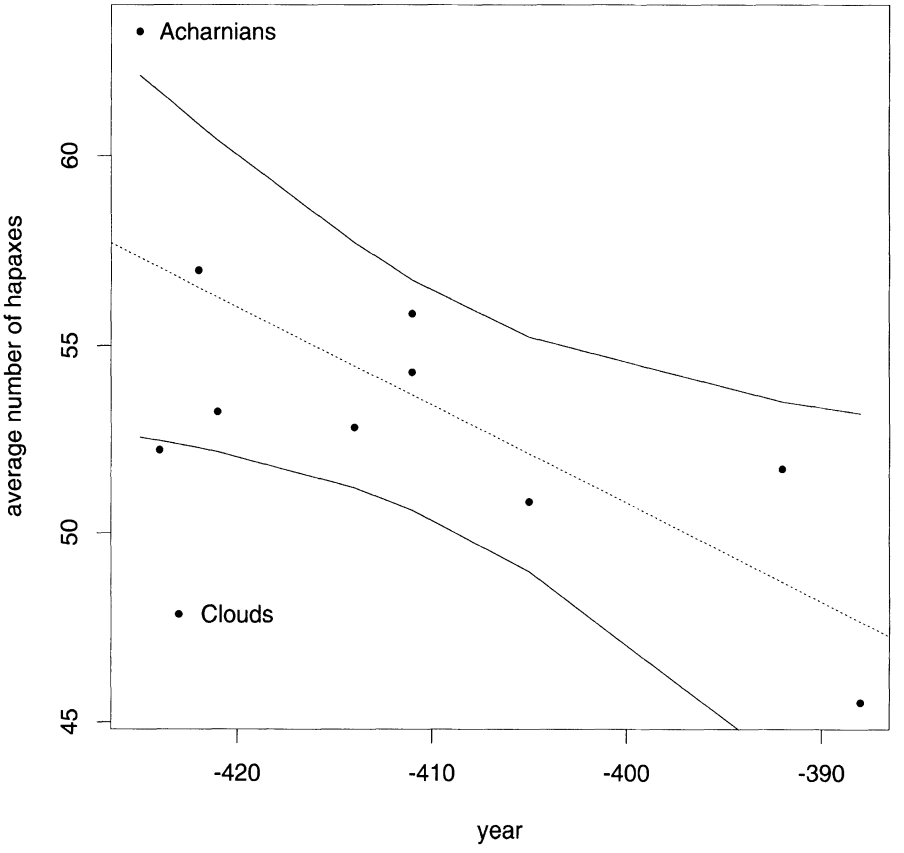

Figure 1. Plot of the average number of *hapaxes* versus date of production for eleven plays of Aristophanes. The regression line (dotted line) was estimated using all plays except *Clouds*. The solid lines represent the 95% confidence bands.

the date of this play. The one-sided 95% inverse prediction interval estimates that *Clouds* was significantly redrafted sometime after 411 B.C.E. (the point estimate indicates that the play was redrafted in 389 B.C.E.). The average vocabulary did not change significantly over time. The average Yule's K increased over time, but not significantly (see Table IV).

Considering each entire play as the block on which to estimate our measures of literary style, we found that the number of words in iambics, N, and Yule's constant changed significantly over time. The number of iambic words per play increased at a rate of about forty-five words per year (see Table IV), which was statistically significant (p-value = 0.003). This regression equation was used to predict the date of *Clouds* to be 415 B.C.E. with a 95% prediction interval indicating that it was redrafted sometime after 424 B.C.E. Yule's constant, K, increased significantly over time (p-value = 0.047), but the large variability in this measure made prediction of the date of *Clouds* impossible, although it is clear that *Clouds*

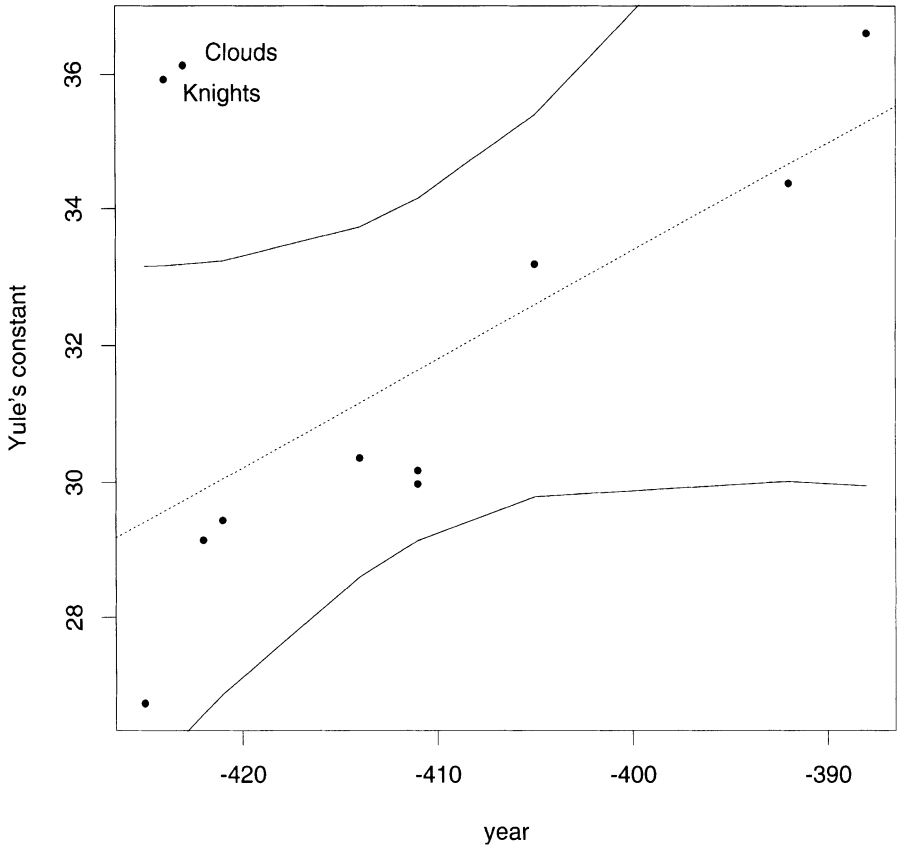

Figure 2. Plot of Yule's constant versus date of production for eleven plays of Aristophanes. The regression line (dotted line) was estimated using all plays except *Clouds*. The solid lines represent the 95% confidence bands.

displays attributes of Aristophanes' later works (see Table IV and Figure 2). Zipf's Z did not change significantly over time.

5.2. TERENCE

The production dates of all six of Terence's plays are known; yet *Hecyra* appears to have been staged in 165 and 160 B.C.E. Table II displays these plays, their performance dates, and the measures of interest. All regression lines were estimated excluding data on *Hecyra*, since the date of this play was uncertain.

The ANOVA analyses of the average number of *hapaxes* and the average Yule's constant indicated that these measures did not vary significantly from play to play (p-values = 0.113 and 0.635). The average vocabulary, however, did appear to vary somewhat from play to play, although this was not significant at the 0.05 level (p-value = 0.071).

Table II. Measures of lexical richness for six plays of Terence

Play	**Date**	$\bar{v} \pm s.e.(v)$	$\bar{h} \pm s.e.(h)$	$\bar{k} \pm s.e.(k)$	K	Z	N
Andria	166	216 ± 2.3	35 ± 1.7	42 ± 2.1	36	45	5966
Hecyra	($\geq$ 165)	218 ± 1.8	35 ± 2.5	38 ± 1.4	34	45	5156
Heaton Tim.	163	217 ± 1.9	36 ± 1.6	40 ± 2.0	35	45	6170
Phormio	161	222 ± 2.1	38 ± 1.8	37 ± 1.6	35	38	6197
Eunuchus	161	222 ± 1.9	34 ± 1.2	40 ± 2.5	33	42	6248
Adelphoe	160	221 ± 1.6	41 ± 2.3	39 ± 2.3	35	36	6741

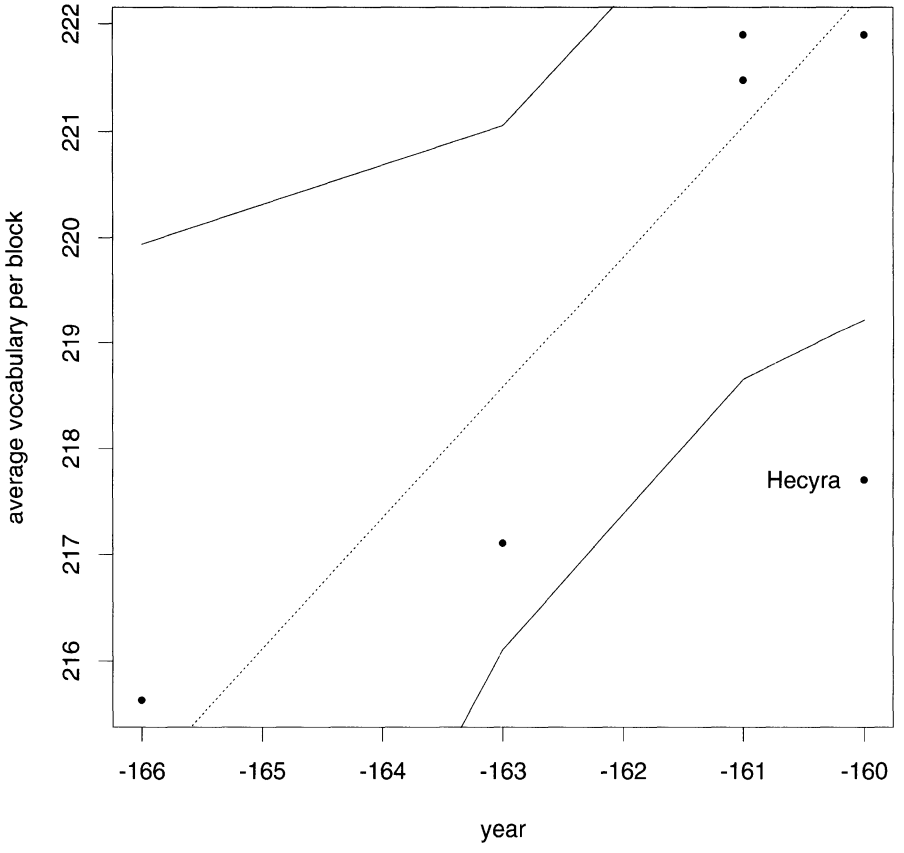

Figure 3. Plot of the average vocabulary versus date of production for six plays of Terence. The regression line (dotted line) was estimated using all plays except *Hecyra*. The solid lines represent the 95% confidence bands.

The weighted linear regression for the average vocabulary was, nevertheless, significant, most likely due to the robustness of this test to heterogeneous variances. The average vocabulary increased over time at a rate of about five words per 300 words per four years (see Table IV and Figure 3). The regression line predicts that *Hecyra* was written in 164 B.C.E. The small number of plays analyzed makes this prediction somewhat uncertain: the 95% prediction interval goes from 172 to 158 B.C.E. Neither the average number of *hapaxes* nor the average Yule's constant changed significantly over time. None of the measures calculated on the whole plays was found to change significantly over time.

5.3. EURIPIDES

Although we have data on nineteen plays of Euripides, only the middle fourteen plays were used in this analysis.[9] Table III displays all nineteen plays and their estimated dates of production. ANOVA analyses found that only the average number of *hapaxes* were significantly different across the plays (Kruskal-Wallis test p-value = 0.034). The p-values for the ANOVA tests for average vocabulary and the average Yule's constant were 0.117 and 0.161 respectively. A weighted linear regression line fit to the *hapax* data reveals that *Supplices* is an outlier to a trend of increasing *hapaxes* over time (see Figure 4). Since *Supplices* is without a certain date, we removed it from the analysis and then found that the average number of *hapaxes* is significantly increasing (p-value = 0.010) over time at a rate of one *hapax* per five years per block of 300 words (see Table IV). *Electra* also appears as somewhat of an outlier in the graph, but was not excluded. The large degree of variability in the data precludes us from estimating the date of either *Supplices* or *Electra*, but it is clear that they each display attributes of later plays. Neither the average vocabulary nor the average Yule's constant changed significantly over time. None of the measures calculated on the whole plays were found to change significantly over time.

6. Discussion

The partitioning method was developed in order to avoid the problems of the text-length dependence of various measures of lexical richness. Using the assumption that words are used randomly and independently in texts, Tweedie and Baayen (1998) showed that both Yule's constant and Zipf's Z are constant across all text lengths. A non-random word usage model showed, however, that these measures changed systematically with text length. Since Yule's constant was calculated both on the whole plays (approximately 5000 words) and on the blocks of 300 words, we can compare these values to determine whether this measure is consistent over different text lengths. In Tables I, II and III, K is consistently about 86% of $\bar{k}$. Tweedie and Baayen (1998) also observed a decrease in Yule's constant with increasing text length under the non-random model of word usage.

Table III. Measures of lexical richness for nineteen plays of Euripides. Production dates marked with daggers are averages of projected dates based on metrical evidence given by Devine and Stephens (1981)

Play	**Date**	$\bar{v} \pm s.e.(v)$	$\bar{h} \pm s.e.(h)$	$\bar{k} \pm s.e.(k)$	K	Z	N
Cyclops	?						
Alcestis	438						
Medea	431	234 ± 2.6	35 ± 3.3	31 ± 1.3	28	42	6512
Hippolytus	428	234 ± 4.4	37 ± 4.6	30 ± 1.9	26	54	6149
Heracleidae	426†	233 ± 2.8	35 ± 2.8	32 ± 2.1	28	49	5419
Andromache	425†	236 ± 4.1	35 ± 2.8	28 ± 1.8	26	49	5866
Hecuba	424	235 ± 2.7	34 ± 2.3	27 ± 1.5	25	52	5799
Supplices	422†	238 ± 3.3	43 ± 3.4	28 ± 2.0	25	52	5631
Electra	416†	236 ± 2.7	42 ± 7.4	30 ± 2.0	26	48	6093
Troades	415	241 ± 2.3	39 ± 3.9	25 ± 1.8	24	58	4817
Heracles	413†	238 ± 1.8	40 ± 3.0	28 ± 2.1	25	52	6045
Iphigenia at Tauris	413†	235 ± 4.1	38 ± 6.5	31 ± 3.7	26	46	6936
Helen	412	235 ± 3.3	37 ± 3.6	30 ± 2.1	27	48	7960
Ion	411†	233 ± 2.4	40 ± 10.7	30 ± 1.3	26	48	6625
Phoenissae	409	237 ± 2.5	39 ± 6.6	30 ± 2.4	26	43	6924
Orestes	408	233 ± 2.7	37 ± 2.6	30 ± 1.8	27	47	6985
Bacchae	405						
Iphigenia at Aulis	405						
Rhesus	?						

Table IV. Weighted Regression Estimates

Author	**Measure**	**slope**	**p-value**	R^2
Aristophanes	$\bar{h}$	–0.26	0.008	0.601
	$\bar{k}$	0.15	0.106	0.293
	N	44.90	0.003	0.695
	K	0.16	0.047	0.406
Terrence	$\bar{v}$	1.23	0.017	0.885
Euripides	$\bar{h}$	0.17	0.103	0.206
(excluding *Supplices*)	$\bar{h}$	0.21	0.011	0.458

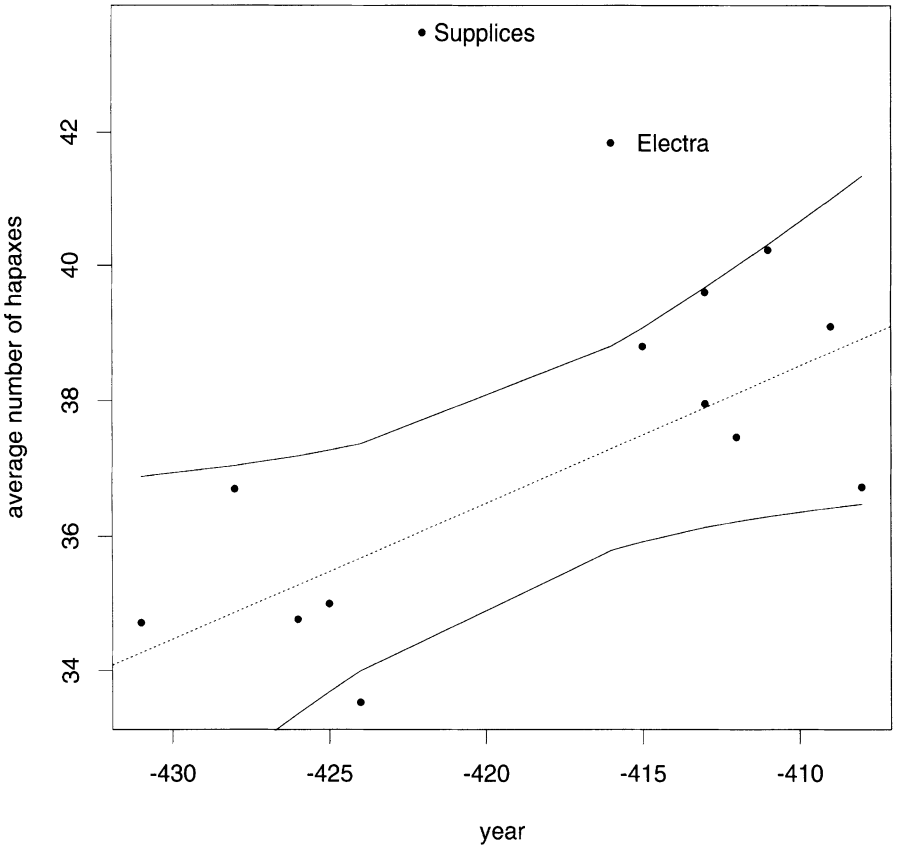

Figure 4. Plot of the average number of *hapaxes* versus date of production for fourteen plays of Euripides. The regression line (dotted line) was estimated using all plays except *Supplices*. The solid lines represent the 95% confidence bands.

We have chosen of a chunk size of $N = 300$ words to use as the standard unit for our expressions of lexical richness based upon two primary considerations. First, given that the spoken, iambic components of the typical classical drama comprise roughly 5,000 to 6,000 words, $N = 300$ allows us to break each play into a sufficient number of samples to derive a fairly stable average per play for our key measures. One may find a playwright employing vocabulary according to the highly particular narrative needs of certain sections of a play (as opposed to the more formulaic sections). Rarely occurring vocabulary will be highly concentrated into those sections. The prologues of plays, sometimes meta-narratively commenting on the content of the plays, tend to bear this distinctive specialized vocabulary. Given a sufficiently small block size, however, these passages of high lexical richness are better isolated into local blocks and the character of less anomalous sections of the text is better observed. Second, $N = 300$ is sufficiently large to observe frequently used vocabulary repeated at rates characteristic of their frequency in the corpus overall. While the vocabulary richness of a 300 word block

is still quite high and dominated by rarely occurring vocabulary (Euripides, for instance, has an average rate of 235.7 types per 300 tokens in his iambics), there is sufficient repetition of frequently used words to make meaningful observations of common types. For example, in a 300 word block from Euripidean trimeter, one may expect the most frequently occurring word to appear, on average, 6.5 times; even the tenth most frequently used word in Euripides will appear, on average, twice (2.2 times), and the twentieth most frequently used word will appear, on average, more than once (1.5 times). We have data suggesting that even a chunk size of 150 words may show some of the signs of repetition of highly favored word types; Euripides, partitioned into 150 word chunks, has a mean vocabulary of 128.0 words, but only four of these words have a better than average chance of appearing twice or more.

Our standard unit will not likely divide evenly into the total number of words of a given corpus. So that we do not end up with some major portion of the standard unit of text left as an uncounted remainder at the end of the last work of a corpus under consideration, we partition so that an equal, minute number of words is dropped out between all chunks. After partitioning we also remove from the data those chunks which span across the end of one work and the beginning of another. Thus we derive a data set for each play which will not describe the complete vocabulary or the total number of *hapaxes* for a given play. But, more importantly, we achieve comparable data sets from each and every work in a corpus. Should it be that an author's signature of style is to pack some portion of a work with particularly dense vocabulary, our method of averaging takes this character into account equally with each work in the corpus.

It becomes imperative, though, that the works compared within a corpus be generically and constitutionally similar: a work of an author that happens to be missing its prologue or conclusion will not likely generate a set of text chunks comparable to a complete play. Such considerations have motivated us to focus our study only on iambic sections of plays, making it possible to compare plays which lack or have greatly diminished lyric sections (e.g. Aristophanes' *Plutus*), to drop from corpora such plays which are clearly generically distinct from other plays in a corpus (e.g. Euripides' satyr and prosatyr plays, *Cyclops* and *Alcestis*), and to drop from consideration those plays too badly mutilated or supplemented by interpolation to be considered an author's complete work (e.g. Euripides' *Bacchae* and *Iphigenia at Aulis*).

6.1. ARISTOPHANES

Figure 2 shows both *Clouds* and *Knights* as outliers to the increasing trend in Yule's constant calculated on the entire plays; the average Yule constant derived from 300 word blocks for these two plays (shown in Table I) is also unusually high compared to other early comedies. And yet the reason for these high average Yule scores for *Clouds* and *Knights* appears to be different in the case of each play.

Knights contains two contiguous blocks of text (representing text from vv. 1009–1126) with unusually large Yule scores, 64.0 and 69.8 (these blocks have z-scores of 3.0 and 3.85 when compared with the rest of the play). Excluding these two blocks and the first block (which has an unusually low Yule's constant) yields a mean and standard deviation of 43.6 ± 6.8 (compared to 45.6 ± 11.9),[10] a value which is still large compared to his other plays of that time. Nevertheless, *Knights* has a unique element in its plot that may explain its deviance from the trend displayed in Figure 2.[11] The play features two roguish figures vying in quasi-political debate to become the chief steward of a household. One of these figures, a sausage-seller by trade, has no natural skill with language and he is prone to echo and mimic the rhetoric he hears from his rival. In addition, the plot calls for an unusual amount of quotation and interpretation of oracular texts to argue who is best suited to be steward. Now such competitions of one-ups-manship are quite familiar to the plots of Aristophanes, but the agonistic rivalry of this play produces repartee inclined to repetition of words, phrases, and sentences, especially in the sections of "oracular" interpretation (e.g. vv. 753–1226 and vv. 116–478). The effect of this duplication of language is reflected in the dearth of infrequent vocabulary. This high rate of recycling of "specialized vocabulary" and the repetition of phrases and sentences may explain the relative lexical poverty in *Knights*.

Clouds, on the other hand, does not have any unusually high average Yule scores due to an anomalous scene of repetitious vocabulary. Rather, its overall high rate of reliance upon frequently used vocabulary is consistent throughout the entire play. The chunk of 300 words with the highest k in *Clouds* is barely one and a half standard deviations above the mean. This sustained usage of common words in *Clouds* at a rate higher than other early comedies signals that its language better reflects the compositional method of Aristophanes late in his career.

Aristophanes composed *Clouds*, saw it lose in a dramatic competition of 423 B.C.E., and then set to work redrafting the play. The extent of this redrafting cannot be known, but it does seem to pervade the whole comedy and involve more than cosmetic touch-ups. There is no record of Aristophanes' re-entry of the redraft and the version we have appears not to have been officially staged again by the comedian (Meineck, 2000; Rosen, 1997). Of interest to the stylometrician is whether *Clouds* still shows stylistic characteristics of Aristophanes' earlier works composed near the time of *Clouds*' production, or whether it shows signs of late redraft. The two measures of vocabulary which we have found significantly correlated to time, the number of words written in iambics per play and the average number of *hapaxes* per standard measure, offer slightly different answers, but with substantial overlap. The date of 415 B.C.E. predicted by the regression equation for the total number of words composed in iambics per play (the 95% one-sided prediction interval has an early limit of 424 B.C.E.) accords with topical references of some of the play's jokes which are historically relevant to the period 419–416 B.C.E. (Meineck, 2000; Rosen, 1997). Yet, the inverse prediction interval based on the tendency in Aristophanes to generate, on average, fewer *hapaxes* as he aged,

predicts an extremely late redraft date for *Clouds* (the 95% one-sided prediction interval has an early limit of 411 B.C.E.). If this prediction is to be believed, then one must think of Aristophanes working on *Clouds* during or after that phase of Athenian democracy when the historic Socrates, satirized and lampooned in *Clouds* as the emblem of the sophistic movement in Athens, sustained popular enmity for his association with Alcibiades (who both sponsored and undermined of the disastrous Sicilian expedition, 415–413 B.C.E.), for his apparent defense of the "anti-democratic" behavior of the generals in the disaster following the victory at Arginusae (406 B.C.E.), and for his perceived sponsorship of the views of Thirty Tyrants who ruled following the fall of the city (403 B.C.E.). Our point estimate date of 389 B.C.E., based on the play's dearth of *hapaxes*, would even have us conceive of the redrafting of *Clouds* through the very last phase of Aristophanes' career. This prediction of extremely late redraft does not preclude earlier redrafting, but does suggest the possibility that the author was still at work on the play following (and, perhaps, because of) the trial and execution of Socrates for his unorthodox views on the gods and for his corruption of Athenian youth (399 B.C.E.). That Plato (*Apology* 18a–d) made an allusion to *Clouds* suggesting that it had contributed to popular misconception about Socrates for some twenty-four years following its staging – a rather large claim considering that the original audience thought so little of it as to inspire Aristophanes' redraft – provides collaborative testimony to a fourth century promulgation of a revised text (when Plato could have accessed the redraft).

6.2. Terence

There is explicit evidence which ties one of Terence's plays, *Hecyra*, to three distinct performance dates (though two of these dates were within the same year, 160 B.C.E.). As in the case of Aristophanes' *Clouds*, Terence sought to have *Hecyra* re-produced due to the poor reception it received in its first showings. Unlike Aristophanes' *Clouds*, however, Terence's multiply-produced play bears only perfunctory indications of redraft: it has two, inorganic prologues. The first prologue (vv. 1–8) mentions a previous, unsuccessful staging. The second prologue (vv. 9–57), composed, presumably, for the occasion of its final successful performance, mentions both a first and a second unsuccessful staging. It is of interest, again, to attempt to determine whether the play carries the stylistic markers of Terence's early compositional technique (i.e. when he first composed the play), or whether it shows "late" improvement through redraft at the end of Terence's career. Unfortunately, Terence's career was wondrously brief, and any stylistic difference between "early" Terence and "late" Terence would be accounting for change over just six years. A regression equation describing Terence's increasing average vocabulary density predicts that *Hecyra* was composed in 164 B.C.E., the year after the play's *didascalia* tells us it was first produced (Ireland, 1990). What credence is to be placed in this prediction is supported by the play's own prologues

which insist that the play needs to be given the fair hearing it deserves because it has yet to be seen by audiences. The statement may be a facetious excuse for re-staging, in 160 B.C.E., a play then five years old, but it tends to verify that *Hecyra*, as we have it, is "early" Terence.

6.3. EURIPIDES

Even when reduced to a fourteen play corpus, Euripides' work proves unwieldy. Only seven of these fourteen plays have production dates firmly fixed either by *didascaliae* or by external criteria. The remaining seven receive less secure production dates based externally on literary referencing (such as by Aristophanic parody), the internal evidence of metrical resolution (Devine and Stephens, 1981), and, most perilously, purported topical allusion within the plays. We therefore chronologically rank these plays more tentatively than those in the corpora of Aristophanes and Terence. That said, the trend in Euripides' average number of *hapaxes* to increase with time is interrupted by the unusually high number of *hapaxes* in the averages for *Supplices* and *Electra* (Figure 4). That neither one of these plays has a firmly fixed production date may appear to invite a prediction of placement somewhere along the regression line at a late stage of Euripides' career. Yet, as with Aristophanes' *Knights* above, the specifics of the data for both *Supplices* and *Electra* may warn against such a prediction. In the case of each play, we find particular, localized scenes with intense *hapax* usage at such a high rate that the mean usage for the entire play is somewhat overestimated.

Electra has one unusual block of text (from vv. 815–872) with seventy-four *hapaxes*, producing a z-score of 3.7 when compared to the rest of the play (mean = 40.0, st.dev. = 9.2). This passage involves the extraneous narrative of a character other than the principals of the play. *Electra* vv. 774–858 feature a messenger describing in long narrative a ritual sacrifice that culminates with the murder of the sacrificer. The whole speech contains *hapaxes* appearing at a rate close to three standard deviations higher than the play's average. The text from vv. 428–518, containing fifty-nine *hapaxes*, derives from two connected scenes in which *Electra*'s famous quasi-comical gentleman peasant has a brief monologue, followed by the arrival of another "minor" peasant figure (487ff.) whose unusually prolonged part in the tragedy has long been suspected as a late interpolation (Kovacs, 1998).

In the case of *Supplices*, we found a scene of some one hundred lines (vv. 634–730) which broke into two chunks of 300 words and rendered sixty *hapaxes* apiece, four standard deviations above the rest of the play's average rate (mean = 41.2, st.dev. = 4.7). The scene again happens to involve the somewhat extraneous material of a messenger's narrative description of a battle taking place off stage. But in this distinctive (perhaps defective) feature, *Electra* and *Supplices* behave quite like three other late Euripidean tragedies: *Iphigenia at Tauris*, *Ion*, and *Phoenissae*. Each of these contains speeches delivered by minor walk-on characters (such as

messengers or attendant slaves) so rich in *hapax legomena* as to be three or more standard deviations above the average for their respective plays.[12] *Electra* and *Supplices*, it seems safe to say, show the characteristics of composition from the later phase of Euripides' career, following *Troades*, as opposed to the earlier phase, prior to *Hecuba*.

7. Conclusion

Though they experienced vastly different careers as writers, Euripides and Terence, as they continued to write, produced plays with richer iambic vocabulary. Aristophanes, on the other hand, drew upon an increasingly reserved palette of word types as he matured and his vocabulary richness, over all, declined with age. If it is true that some aspects of the style of these authors changed over time for reasons other than those connected with the internal process of poetic craftsmanship, if fluctuations in the social climate and theatrical traditions of the time influenced authorial style, then it is interesting to observe that the same set of environmental factors could produce the opposite results in the compositional style of the contemporaneous careers of the tragedian Euripides and the comedian Aristophanes. And while it is safe to say that the vocabulary richness of the small chunks of texts that we examined is most powerfully affected by an author's ongoing generation of new and rare word types, one of our three authors has provided us with evidence that the increase in rare word usage and vocabulary richness overall is no sure sign of increase in maturity and ability.

Our exploratory studies in the characterization of vocabulary richness by averaging partitioned chunks give sufficient indication that it is possible to correlate this aspect of authorial style with chronology. Thus our partitioning method may join the ranks of other established stylometric chronometers: the relative frequency of common function words, the frequency of metrical anomalies (in poetic corpora) such as resolved feet, and (again, in poetry) the pattern of agreement between word boundary or sense pause and metrical position, to name the most common (Fitch, 1981; Frischer, 1991; Laan, 1995). We make no representation that our various measures characterizing average lexical richness, when found to be significantly changing with time across literary corpora, offer a more powerful or dependable chronometric gauge than these others for the plotting of a work of unknown date against known works. The stylometrician will wish to brace the results offered by this data with other available measures. And yet in surmounting the traditional obstacles to the holistic characterization of an author's literary vocabulary, we gain a clear sense of the extent to which compositional style, even within the highly restricted parameters of composition here observed, is unstable.

Notes

[1] These are the seven tragedies ascribed to Aeschylus (active, c. 498–c. 456 B.C.E.), the seven tragedies of Sophocles (active, 468–c. 405 B.C.E.), the nineteen plays ascribed to Euripides (active, 455–c. 406 B.C.E.), the eleven comedies of Aristophanes (active, 427–c. 387 B.C.E.), the four nearly complete comedies and abundant fragmentary comedies of Menander (active, c. 322–c. 291 B.C.E.), the twenty-one comedies ascribed to Plautus (active, c. 215–186 B.C.E.), the six comedies of Terence (active, 166–160 B.C.E.), and the ten tragedies ascribed to Seneca (active ?37 C.E.–?65 C.E.).

[2] There are exceptional instances of authorial redraft, e.g. Aristophanes' *Clouds* or (perhaps) Terence's *Hecyra*.

[3] Tradition ascribes at least seventy plays to Aeschylus' forty year career, eighty-eight or more to Euripides' fifty years of writing, and over 130 comedies to Plautus' thirty year career. Of our eight classical playwrights, Terence, who composed only six plays, is the exception; but even he composed these six plays in a brief, eight year career.

[4] Six of the seven plays of Aeschylus can be securely dated by the external evidence of *didascaliae*, but that corpus is here excluded from consideration on the grounds that (i) one of the seven plays is of doubtful authenticity, (ii) another of the seven, a history play, is generically distinct from the tragedies, and (iii) three of the remaining five plays were written as part of a tetrology for performance on the same day. And so with only three chronological data points shared among five tragedies, we judge Aeschylus' corpus to be unsuited to the type analysis pursued here.

[5] Calculating this measure on chunks of text much smaller than a whole play is likely to be inaccurate and have substantial estimation error.

[6] Details of the techniques used in preparing the texts may be obtained by application to the authors.

[7] Note that the number of chunks per play, n, differs as the size of plays may substantially differ. Our rationale for the selection of $N = 300$ words per chunk is explained below.

[8] Baayen (1991) in his study on the generation of new words in the context of natural language (i.e. non-literary), maintains the importance of characterizing generative productivity in terms of the ratio of *hapaxes* to the whole sample as this best predicts the likelihood of further new coinages as sample size increases. Baayen *et al.* (1996) have further argued that, in a literary context, the measure of *syntactic hapaxes* (i.e. sentence structures occurring once in a sample) indexes "an author's syntactic creativity, and can be used to gauge how well an author has mastered the possibilities offered by the grammar". Given that we are observing word forms being generated entirely within the limiting conventions and traditional rules of metric poetry in drama, our $\bar{h}$ may indicate the mastery of and creative modifications to those rules of poetic composition. But see, however, the trend in generation of *hapaxes* over time in the plays of Aristophanes, Figure 1 below.

[9] Excluded from consideration were: *Rhesus*, because of its doubtful authenticity, *Cyclops* and *Alcestis*, because these plays, one a satyr play and the other a prosatyr play, are generically different from Euripides' tragedies, and *Bacchae* and *Iphigenia at Aulis*, because these plays, performed not until after the death of Euripides, seem to have suffered an unusual amount of actor interpolation and contamination. *Bacchae* is lacking considerable portions of its conclusion. On this and the problem of the pervasive interpolation in *Iphigenia at Aulis*, see the critical appatus of Diggle (1994).

[10] These values have been rounded in Table I, where the standard error is listed rather than the standard deviation.

[11] In addition to *Knights* vv. 1078–1180 ($k = 69.8$) and vv. 1009–1077 ($k = 64.0$) high Yule scores are found in vv. 163–211 ($k = 52.3$) and vv. 211–478 ($k = 50.0$).

[12] *Ion* vv. 1126–1263 broke into two contiguous chunks which contributed ninety-two and sixty-seven *hapaxes* apiece, yielding z-scores of 8.5 and 4.7 relative to the rest of the play; the 300 words of *Phoenissae* vv. 1098–1152 produced seventy-four *hapaxes* (z-score of 3.9); *Iphigenia at Tauris* vv. 1356–1407 produced sixty-five *hapaxes* (z-score of 3.2), and *Iphigenia at Taurus* vv. 242–295 produced sixty-two *hapaxes* (z-score of 2.9).

References

Baayen, H. "Quantitative Aspects of Morphological Productivity". *Yearbook of Morphology*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991, pp. 109–49.

Baayen, H., H. Van Halteren and F. Tweedie. "Outside the Cave of Shadows: Using Syntatic Annotation to Enhance Authorship Attribution". *Literary and Linguistic Computing*, 11 (1996), pp. 121–131.

Cropp, M. and G. Fick. *Resolutions and Chronology in Euripides: The Fragmentary Tragedies*, Institute of Classical Studies, London, 1985.

Devine, A. and L. Stephens. "A New Aspect of the Evolution of the Trimeter in Euripides". *Transactions of the American Philological Association*, 111 (1981), pp. 45–64.

Diggle, J. *Euripides: Fabulae Tom. III*, Clarendon Press, Oxford, 1994.

Draper, N. and H. Smith. *Applied Regression Analysis, Second Edition*, John Wiley and Sons, New York, NY, 1981.

Duckworth, G. *The Nature of Roman Comedy*, Princeton University Press, Princeton, NJ, 1952.

Fitch, J. "Sense-Pauses and Relative Dating in Seneca, Sophocles and Shakespeare". *American Journal of Philology*, 102(3) (1981), pp. 289–307.

Frischer, B. *Shifting Paradigms: New Approaches to Horace's Ars Poetica*, Clarendon Press, Oxford, 1991.

Ireland, S. *Terence: The Mother-In-Law*, Aris and Phillips Ltd., Warminster, 1990.

Kovacs, D. *Euripides: Vol. III*, Harvard University Press, Cambridge, MA, 1998.

Laan, N.M. "Stylometry and Method. The Case of Euripides". *Literary and Linguistic Computing*, 10(4) (1995), pp. 271–278.

Meineck, P. *Aristophanes: Clouds*, Hackett Publishing Co., Indianapolis, IN, 2000.

Orlov, J.K. "Ein Model der Häufigkeitsstruktur des Vokabulars". *Studies on Zipf's Law*, Brockmeyer, Bochum, 1983, pp. 154–233.

Rosen, R.M. "Performance and Textuality in Aristophanes' *Clouds*". *Yale Journal of Criticism*, 10(2) (1997), pp. 397–421.

Tweedie, F. and H. Baayen. "How Variable May a Constant be? Measures of Lexical Richness in Perspective". *Computers and the Humanities*, 32 (1998), pp. 323–352.

Yule, G. *The Statistical Study of Literary Vocabulary*, Cambridge University Press, 1944.