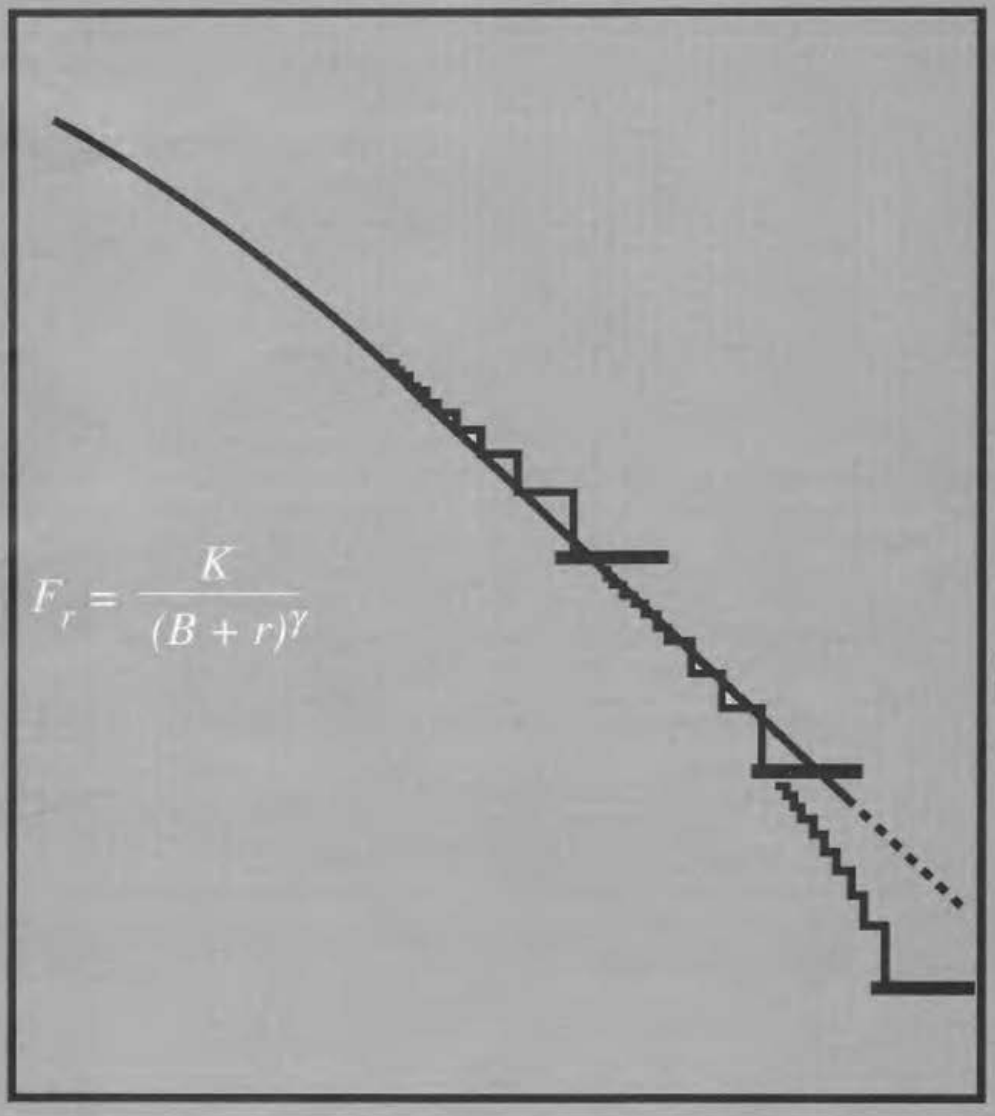
Reprint form:

ISSN 0929-6174
CODEN JQLIEO

Journal of Quantitative Linguistics



SWETS & ZEITLINGER
PUBLISHERS

Journal of Quantitative Linguistics
1999, Vol. 6, No. 1, pp. 85–97

0929-6174/99/0601-0085$15.00

Analysis of Classical Greek and Latin Compositional Word-Order Data*

Fiona J. Tweedie[1] and Bernard D. Frischer[2]
[1]University of Glasgow, United Kingdom
[2]University of California, Los Angeles, USA

ABSTRACT

A recent paper by Frischer et al. (forthcoming) examines the position of the direct object and its governing verb in works in Classical Greek and Latin. The paper confirms the SOV ordering expected in texts in Latin, and the S(O)V(O) ordering expected in texts written in Greek. Texts written in Greek by Cassius Dio were found to have a Latinate word-order. However, subsequent research on the statistical analysis used in that paper has resulted in a refinement that will be presented in the present paper. For each of the sixty texts examined, one hundred direct objects were categorised. The knowledge of the total number of direct objects examined constrains the data; it is described as being compositional. In this paper we re-examine the data using techniques developed for compositional data analysis. The conclusions of Frischer et al. are confirmed and new insights into texts by Marcus Aurelius and Plutarch are obtained.

INTRODUCTION

A recent paper by Frischer et al. (forthcoming) examines the position of the direct object and its governing verb in works in Classical Greek and Latin. Through statistical analysis of a great number of sentences, one long suspected difference between Latin and Greek word-order was confirmed, and the ramifications of this observation were explored for some possible cases of word-order transference between Latin and Greek. The difference between the languages concerns the positioning of the accusative direct object with respect to the verb governing it. That there is a difference in the Greek and Latin distributions is no surprise: Classical linguists have long observed that Latin has a greater tendency to place the verb at the end of the clause than does Greek. From this fact alone one might predict that the direct object in Latin is more likely to precede than to follow the verb on which it depends than is the case in Greek. This prediction was tested empirically by tabulating the direct object distributions in sixty passages written by fifteen Latin and ten Greek prose authors. Each passage was randomly selected in the text of an author. Analysis was based on the first one hundred direct objects in the accusative case that were encountered in a passage, and a tabulation was done of those that occurred before and those after the governing verbs. With remarkable consistency, the texts in our sample clumped into a Latin and a Greek cluster, offering strong empirical and statistically significant proof that the position of the direct object with respect to its governing verb differed in Greek and Latin prose.

Five Greek texts by two authors writing in Greek turned out to be anomalous, fitting firmly into the Latin group. Four of the texts were writ-

*We are grateful to Dr J. W. Kay of Glasgow University for helpful advice as well as the code for the permutation test of equality of covariances based on Box's *M* statistic. We would also like to thank Professor J. Aitchison of Glasgow University for his support and helpful comments on a draft of this paper.

Address correspondence to: Fiona J. Tweedie, Department of Statistics, University of Glasgow, University Gardens, Glasgow, G12 8QW, United Kingdom.

ten by Cassius Dio; the fifth is the Greek translation of the Emperor Augustus' *Res Gestae*, a bilingual version of which survives, and whose original is known to have been written in Latin.

In considering the results, including the anomalous texts, the study showed that native language did not necessarily have an effect on a writer's placement of the direct object; nor did the language of an important literary or historical source. Native Greek authors writing in Latin respected Latin word order; Romans writing in Greek generally conformed to Greek practice.

The study suggested that some but not all explanations for the data are linguistic. On the linguistic level, it was the greater consistency of Latin SOV word-order that helped the Latin pattern to prevail over the more flexible Greek positioning of the verb and direct object. This was true not only for Roman authors writing Latin with a Greek source before them (like Aulus Gellius or Cicero) but also for a Greek author like Ammianus Marcellinus writing in Latin. It was evidently normally easy for both Greeks and Romans to recognize and to respect the tendency of Latin to place the verb at the end of the clause. On the other hand, in the interesting case of the Greek translation of the *Res Gestae* and other official documents, where the Roman chancellery's habit of translating Latin into Greek through quasi-relexification was seen, the study proposed an explanation based either on Roman scrupulosity in legal matters or on a sociological factor of linguistic hegemony. Finally, in the case of Cassius Dio there was seen the operation of a psycholinguistic or sociolinguistic cause for word-order transference: Dio's conscious or unconscious presentation of himself as a Roman.

The data sets used by Frischer et al. are enumerated in the Appendix. They list for each text sample the number of direct objects that occur *before* the governing verb in main clauses (MCB) and other clauses (OB) as well as the number of direct objects that occur *after* the governing verb in main clauses (MCA) and other clauses (OA). In most cases the total number of direct objects is 100. There are a few samples where only 99 sentences were examined, the data have been rescaled to have a total of 100. Frischer et al. (forthcoming) treat the number of clauses in each of the categories as separate, independent random variables. Quantitative linguists and stylometricians may not be aware of a well-known problem that affects the "standard" statistical analysis of compositional data, that is, when dealing with data that adds up to a known total. This paper aims to explain the problem, present one approach that succeeds in solving it; and apply this approach to such statistical techniques such as principal components analysis, cluster analysis and discriminant analysis. The conclusions of Frischer et al. (forthcoming) will be re-examined.

IT ALL ADDS UP TO 1 – CONSTRAINTS ON THE COVARIANCE MATRIX

The statistical analysis detailed in the previous section produces easily interpretable results; separate clusters of Latin and Greek authors. However, the authors ignore a constraint on the data, that the number of clauses examined is always one hundred. Rather than having four independent random variables, then, only three are ever of interest, the fourth is simply one hundred minus the sum of the other three. Turning to mathematical notation, we define x_{ijk} as the proportion of a particular position of the direct object in language i, $i = G, L$, text number j, $j = 1, \ldots, N_i$, where N_i is the number of texts sampled from language i. The third subscript, $k = 1, \ldots, D$, where $D = 4$, categorises where the direct object occurs in relation to its governing verb; before it (MCB and OB) or after it (MCA and OA). The value of the fourth random variable can be calculated as:

$$x_{ijD} = 1 - \sum_{k=1}^{D-1} x_{ijk}. \qquad (1)$$

This constraint has far-reaching implications, in particular with respect to the covariance structure of the data.

Covariance Structure

Define $\text{var}(x_k)$, $\text{cov}(x_k, x_{k'})$ and $\text{corr}(x_k, x_{k'})$ to be the variance of x_k, the covariance of x_k and $x_{k'}$ and the correlation between x_k and $x_{k'}$ respectively. Then it can be shown that

$$\text{var}(x_k) = \text{cov}(x_k, x_k) \tag{2}$$

for $k = 1, \ldots, D$, and

$$\text{corr}(x_k, x_{k'}) = \frac{\text{cov}(x_k, x_{k'})}{\sqrt{\text{var}(x_k)\ \text{var}(x_{k'})}} \tag{3}$$

for $k, k' = 1 \ldots D$.

We can now define the *crude covariance structure* of a composition x to be the set of all

$$\kappa_{kk'} = \text{cov}(x_k, x_{k'}) \tag{4}$$

for $k, k' = 1 \ldots D$ with the $D \times D$ *crude covariance matrix*

$$K = [\kappa_{kk'} : k, k' = 1 \ldots D], \tag{5}$$

and *crude correlations*

$$\rho_{kk'} = \frac{\kappa_{kk'}}{\sqrt{\kappa_{kk}\kappa_{k'k'}}}. \tag{6}$$

Investigation of the crude covariance structure generally concentrates on properties of the crude covariance matrix K. The patterns of variability for our Greek data can be estimated as

$$\widehat{\kappa_{Gkk'}} = \frac{1}{N_G - 1} \sum_{j=1}^{N_G} (x_{Gjk} - \overline{x_{Gk}})\,(x_{Gjk'} - \overline{x_{Gk'}}) \tag{7}$$

for $k, k' = 1 \ldots D$ and where

$$\overline{x_{Gk}} = \frac{1}{N_G} \sum_{j=1}^{N_G} x_{Gjk}. \tag{8}$$

Values of $\widehat{\kappa_{Lkk'}}$ and $\overline{x_{Lk}}$ for the Latin data are defined in a similar way.

Given a composition x_{ij} which is subject to the constraint:

$$x_{ij1} + \ldots + x_{ijD} = 1, \tag{9}$$

the problems with interpreting the crude covariance matrix are discussed by Aitchison (1986), section 3.3, and are detailed in the following paragraphs.

Negative Bias Difficulty

Since

$$\text{cov}(x_1, x_1 + \ldots + x_D) = 0, \tag{10}$$

we have that

$$\text{cov}(x_1, x_2) + \ldots + \text{cov}(x_1, x_D) = -\text{var}(x_1). \tag{11}$$

The right hand side of this equation is negative, except for the trivial case where x_1 is constant. Thus at least one of the components of the left hand side must be negative, or equivalently one of the elements in the first row of K must be negative. This holds for all D rows, thus there must be at least D negative elements in K. Correlations which are free to range over $[-1, 1]$ with unrestricted data, are constrained. In particular, if $D = 2$ then $\rho_{12} = -1$.

This negative bias that is present in the crude covariance matrix forces the interpretation of covariances to be rather different from the standard interpretation of covariances between components of an unrestricted vector.

Subcomposition difficulty

In some cases we might be interested in a *subcomposition*, where only certain elements of the data are of interest, such as looking only at main clauses. If the full composition is known then the subcomposition can be calculated, but there is no simple relationship between the crude covariance matrix of the subcomposition and that of the full composition. Indeed, the covariance between elements of the composition can change erratically as different subcompositions are examined.

Null correlation difficulty

As shown above, elements in a crude correlation matrix have a negative bias. However, the standard interpretation of covariances and correla-

tions is that zero indicates no association or independence. This interpretation is rather suspect under the negative bias property. Some researchers (for example Chayes & Kruskal, 1966) have proposed techniques that result in *null* correlations, values of the correlation coefficient that indicate a lack of association, that are not necessarily zero. Aitchison (1986, p. 57), lists a number of objections to the method.

Conclusions

It is clear that interpretation of the crude covariance matrix of data with a constrained sum is fraught with problems. Thus any analysis of the data which involves the covariance matrix is also suspect. This includes principal components analysis, discriminant analysis and certain versions of cluster analysis.

COMPOSITIONAL DATA ANALYSIS

Aitchison (1983) notes that various researchers have tried to deal with the problems raised in the previous section by considering alternative covariance matrices.

The first is the use of the covariance matrix of the data, omitting one of the components, say component k, and written as $\mathrm{cov}(x_{-k})$. Aitchison writes that this

> involves a common, though naïve, approach to the analysis of compositional data through the omission of one of the proportions, presumably in the mistaken belief that the remaining proportions are relatively unrestricted. The effect on the interpretation of correlations persists through the inequality $\sum_{k' \neq k} x_{ijk} < 1$ (pp. 59–60).

The second covariance matrix is simply K, the crude covariance matrix. K is singular and of order D and thus has a zero eigenvalue. It has been suggested (Le Maitre, 1968) that the other $D-1$ eigenvectors could be used to form principal components. These principal components are linear combinations of the raw data and are still subject to the sum constraint and correlation problems as outlined above.

The final technique listed by Aitchison is the covariance matrix of the variables (x_{-k}/x_k). While this is better than the two attempts above, principal components derived from it are linear, and dependent on the choice of k.

Aitchison (1983) proposes an alternative to these transformations, the *logratio covariance matrix*:

$$\Sigma_k = \mathrm{cov}\{\log (x_{-k}/x_k)\}. \qquad (12)$$

The use of the logarithmic function allows for the non-linearity often found in compositional data. However, this may still be dependent on the choice of k. Aitchison (1986, section 5.5), shows that for most mathematical operations, the analysis is invariant to the choice of k.

Principal Components Analysis

Aitchison (1983) shows that principal components derived from Σ_k have the form

$$\sum_{l \neq k} \alpha_l \log (x_{ijl}/x_{ijk}) \qquad (13)$$

which, as $\sum_{l=1}^{D} \alpha_l = 0$, can be expressed as

$$\sum_{l=1}^{D} \alpha_l \log (x_{ijl}) \qquad (14)$$

i.e., the principal components can be expressed as log linear contrasts of the crude data. However, these are still subject to the sum constraint. This can be dealt with by noting that, with $\sum_{l=1}^{D} \alpha_l = 0$,

$$\sum_{l=1}^{D} \alpha_l \log (x_{ijl}) = \sum_{l=1}^{D} \alpha_l \log (x_{ijl}/\dot{x}_{ij}), \qquad (15)$$

where $\dot{x}_{ij}$ is the geometric mean, $\dot{x}_{ij} = (x_{ij1} \ldots x_{ijD})^{1/D}$. Thus logcontrast principal components analysis is carried out on the logratio transformed data, $\log(x_{ijk}/\dot{x}_{ij})$, using the *centred logratio covariance matrix*,

$$\Gamma = [\gamma_{kk'}] = [\mathrm{cov}\{\log(x_k/\dot{x}), \log(x_{k'}/\dot{x})\}], \qquad (16)$$

for $k, k' = 1, \ldots, D$, thus removing the need to choose a denominator for Σ_k. Aitchison (1986,

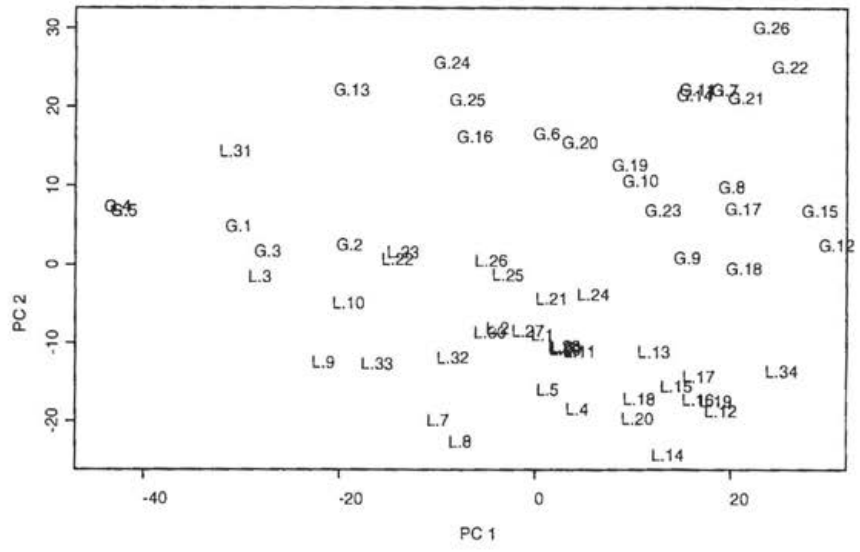


Fig 1. Text samples displayed in the area spanned by the first two principal components.

section 8.3), demonstrates the equivalence of the properties of this logcontrast principal components analysis with the standard principal components analysis for unconstrained data.

We shall proceed to analyse the word-order data in this way. It should be noted, meanwhile, that two of the texts (Caesar's *Bellum Gallicum* Book 3 (L.8), and Cassius Dio's Book 40 (G.3)) have no occurrences of a direct object after a verb in a non-main clause. Thus it would be impossible to carry out a log-transform of these texts. We have dealt with this by changing the zero values to 0.5 and by re-scaling each affected row.[1]

Principal components analysis of the crude data

Figure 1 shows the text samples in the space defined by the first two principal components of the crude data. These components describe 90.36% of the variation in the data. It can be seen that the texts by Latin authors are positioned in the lower half of the graph, and texts by Greek authors are in general in the upper right quadrant of the graph. The first principal component is almost entirely negatively associated with direct objects occurring before the verb in main clauses (MCB, the correlation coefficient between this and the first principal component is –0.99), and positively associated with direct objects occurring after the verb in other clauses (OA, the correlation coefficient is 0.71). The second principal component is negatively associated with direct objects occurring before the verb in other clauses (OB and correlation coefficient 0.89), and positively associated with direct objects occurring after the verb in main and other clauses (MCA and OA, correlation coefficients 0.69 and 0.58 respectively).

It can be seen from Figure 1 then, that the Greek texts in general have a high rate of direct objects appearing after the verb in both main and other clauses. Exceptions to this are the Greek texts at the centre-left of the graph, and the Latin text in the high left. The Greek texts in the centre left are those by Cassius Dio, already noted by Frischer et al. (forthcoming) as being unusual in their word-ordering. The translation

1. Other constants, 0.10, 0.25, 0.75 and 1 were examined. All of the results in this paper are robust with respect to these changes in the constant. The value of 0.5 was chosen as it did not result in the texts with zero points being either too extreme or too close to the other data in the principal components and cluster analyses. Other techniques for dealing with zeroes in the data include the amalgamation of categories, omission of records with a zero category and the use of the Box-Cox family of transforms rather than the log-transform used here. More details can be found in Aitchison (1986) chapters 11 and 13.

into Greek of Augustus' *Res Gestae* (G.1) is also in this area, close to the Latin original (L.3). Frischer et al. note that the word-order in the Greek translation "almost exactly mirrors that of the Latin original" (11). The Latin text that appears in the high left of the graph is Tacitus' *Agricola* (L.31). This text has the second highest number of direct objects occurring after the verb in other clauses in Latin texts, as well as a high number of direct objects occurring after verbs in main clauses (OA).

While there appears to be a clear division between Latin and Greek authors, there is also some curvature present in the graph; there are no texts with low values of the second principal component at the extremes of the first principal component, and there appear to be few texts with high values of the second principal component and values around zero on the first. This curvature is symptomatic of the sum constraint on the data.

Logcontrast principal components analysis

Figure 2 shows the text samples in the space defined by the first two logcontrast principal components. These components describe 89.67% of the variation in the data. It can be seen that the texts by Greek authors are generally on the right of the graph, while the Latin texts are in the lower left area. The first principal component is strongly positively correlated with high values of direct objects occurring after the verb in non-main clauses (OA and a correlation coefficient of 0.90), and negatively associated with direct objects occurring before the verb in main clauses (MCB and correlation coefficient of –0.75). The second principal component is positively associated with of direct objects occurring after the verb in main clauses (MCA and 0.82) and negatively associated with direct objects occurring before the verb in non-main clauses (OB and –0.78). It is thus clear that the Greek authors use direct objects after verbs, in both main and other clauses (MCA and OA), more frequently than Latin authors.

There are some exceptions to the left-right, Latin-Greek divide found in Figure 2. The Greek texts in the centre-left (G.2, G.3, G.4 and G.5) are again those of Cassius Dio, and the Greek translation of the *Res Gestae* (G.1) remains close to its Latin original (L.3). Tacitus' *Agricola* is the Latin text (L.31) that occurs closest to the top right corner. Surprisingly, the other Tacitus texts in this study, the *Annales* (L.32) and *Historiae* (L.33) are to be found in the low centre and far left of the graph, with 89% and 93% of direct

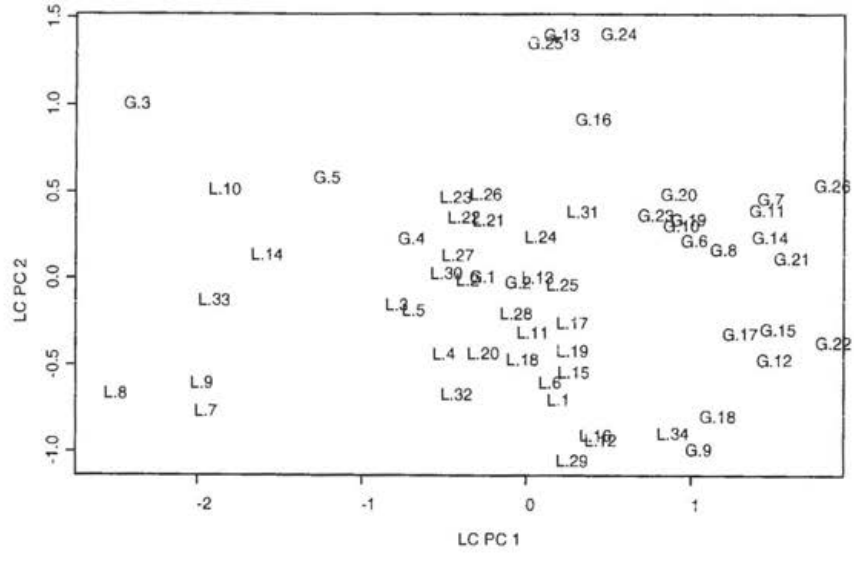


Fig. 2. Text samples displayed in the area spanned by the first two logcontrast principal components.

objects occurring before the verb (MCB and OB) respectively, in comparison with 78.58% for the *Agricola*. One possible solution for this may be the genres in which Tacitus was writing: the *Agricola* is a biography, whereas the other two texts are histories. Another interesting point that was not observed in the crude analysis is the positioning of Varro's *Res Rusticae* (L.34). This is the Latin text that can be found in the lower right of the figure, next to two Greek texts.

These results broadly confirm the results from the crude analysis and add refinements; the perhaps Graecian nature of Varro's word-ordering and the highlighting of genre differences with Tacitus appear to merit further investigation.

Cluster Analysis

The standard form of cluster analysis for unconstrained data requires a measure of distance between two points. The clustering is then performed by an algorithm applied to these distances, for example, by single, average or complete linkage, where the distance between groups is defined as the shortest, average or longest distance between points in the two groups, respectively.

In most cases the squared Euclidean metric,

$$d^2_{ij,i'j'} = \sum_{k=1}^{D} (x_{ijk} - x_{i'j'k})^2, \qquad (17)$$

is used to measure distance between points x_{ij} and $x_{i'j'}$, although others such as the Mahalanobis distance can be used.

For compositional data, however, we need to transform the Euclidean metric. Aitchison (1983, p. 64) proposes the following:

$$d^2_{ij,i'j'} = \sum_{k=1}^{D} [\log (x_{ijk}/\dot{x}_{ij}) - \log (x_{i'j'k}/\dot{x}_{i'j'})]^2 \qquad (18)$$

and we shall analyse the word-order data using this distance metric.

Figure 3 shows the dendrogram resulting from a cluster analysis on the crude data. It is clear that the data divides into four main clusters; from left to right, a Greek cluster (G.13, G.16, G.24 and G.25) a Latin cluster that includes three texts by Cassius Dio (G.3, G.4, and G.5), Tacitus' *Agricola* (L.31), as well as both versions of the *Res Gestae* (L.3, and G.1). Next follows another cluster of Greek texts, and another one of Latin texts. The texts found in the smaller Latin cluster can be described as having

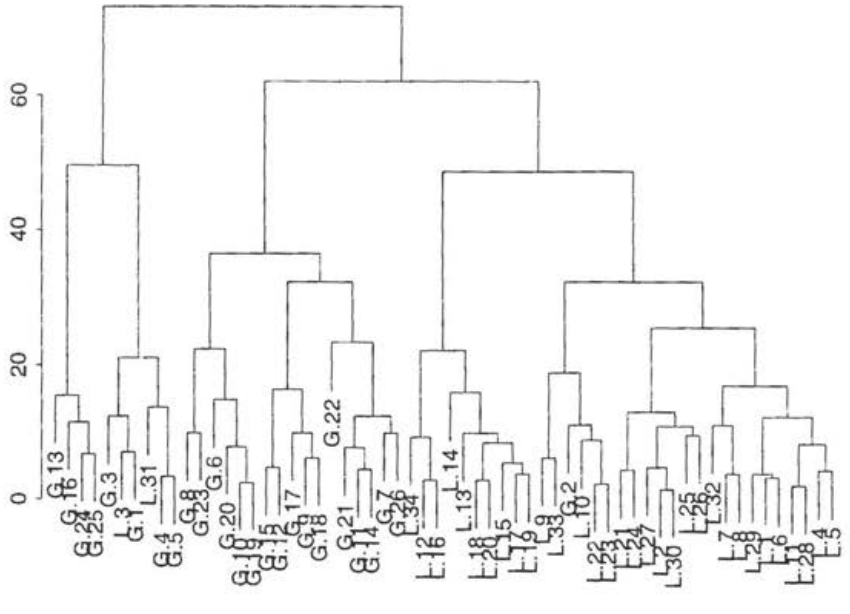


Fig. 3. Dendrogram resulting from cluster analysis of the crude word-order data.

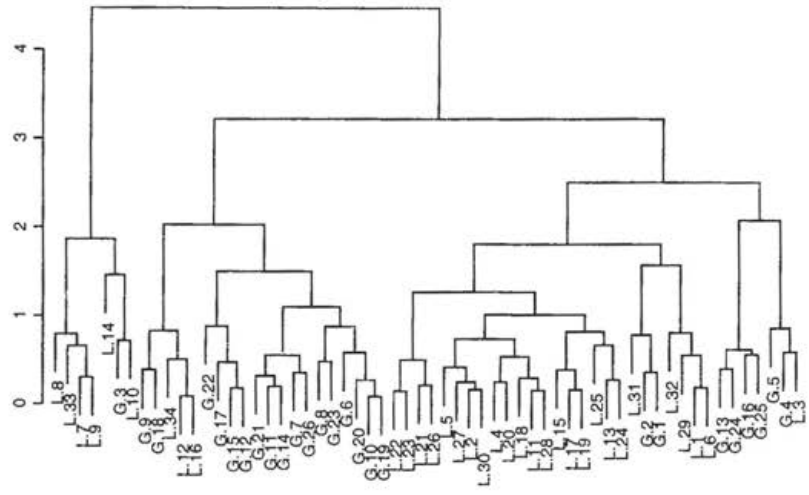


Fig. 4. The dendrogram resulting from revised cluster analysis of the word-order data..

more than double the number of direct objects occurring before the verb in main clauses (MCB), as opposed to other syntactic structures. The Greek text present in the second, larger, Latin cluster (G.2) is Cassius Dio's Book 36, already noted as appearing to have a Latinate word-order.

Figure 4 shows the dendrogram resulting from a cluster analysis using the revised distance metric. Again it can be seen that the texts split into clusters roughly by language, as follows:

1. Latin: L.8, L.33, L.7 and L.9.
2. Latin: L.14, G.3 and L.10; G.3 is Cassius Dio's Book 40, already found to have a Latinate word order.
3. Mixed: G.9, G.18, L.34, L.12 and L.16.
4. Greek: G.22 – G.19.
5. Latin: L.22 – L.6, including G.2, Cassius Dio's Book 36 and G.1, the Greek translation of Augustus' *Res Gestae*.
6. Greek: G.13, G.24, G.16 and G.25.
7. Latin: G.5, G.4 and L.3; G.5 and G.4 are Cassius Dio's Books 59 and 54 respectively, while L.3 is the original version of Augustus' *Res Gestae*.

The first two Latin clusters are the least similar to the remaining texts. They are made up of texts with either a single example or no examples of a direct object occurring after its governing verb in a non-main clause (OA). The nature of the log-transform accentuates this difference, thus texts with this kind of data structure will appear more similar to each other than to any other text.

With the exception of cluster 3, the other clusters clump texts seen together before. Not so those in cluster 3. The original article had in vain sought evidence that source texts in one language could exert a stylistic influence on target texts in the other language; or that native language might affect writing in a second language. Here the corrected metric may at last provide some examples, which would not be at all surprising in view of the previous applied linguistic research on other languages (cf. Odlin, 1990, pp. 98–104; Selinker, 1969).

Cluster 3 includes two Greek texts, Book 3 of the *Meditations* of Marcus Aurelius and Plutarch's *Life of Sertorius*, as well as three Latin texts, Varro's *Res Rusticae*, Cicero's *Brutus*, and Cicero's *De Officiis* Book 1b. Common to the Greek texts is a relatively high sum of MCB and OB; indeed, the sums are the highest among the texts studied by their authors. Marcus Aurelius was a native speaker of Latin who wrote the *Meditations* in Greek, in which he was clearly highly proficient. It has been noted that even

advanced students of a second language may persist in borrowing features from their native language that are not present (or, in our case, we should say that are not as frequent) as they are in their mother tongue (cf. Corder, 1983, p. 95; Kellerman, 1983, p. 114). Although Book 8 of the *Meditations* conforms to good Greek style with respect to the positioning of direct objects, Book 3 appears to betray a more Latinate Sprachgefühl. Book 3 is one of four books that are much shorter than the rest (cf. Brunt, 1974, p. 18) and was probably written in the field at Carnuntum (as is stated in an entry preceding the beginning of the book in ms. P; on the date of composition of Book 3 see Brunt, 1974, pp. 18–19). Whether owing to hasty composition, the absence of Hellenophones at military headquarters on the frontier, or a lack of opportunity to revise and polish (of which there are other signs in the *Meditations*; see Brunt, 1974, p. 5, Aurelius' style in Book 3 may well reflect transference of a feature from his native to his acquired language.

Plutarch raises a different matter. Norden noted that Plutarch's style often varies with his source (Norden, 1958, p. 395), and we may note in this regard that Plutarch's texts fall into more clusters (three) than do those of any other author in our study. It has been suggested that Plutarch composed his biographies by using one source per section (see Pelling, 1988, p. 31), and our data may add plausibility to this theory. Plutarch's sources for the life of Sertorius were Roman (on Plutarch's sources in the *Lives* generally, see Christ, 1905, pp. 679–680. If Plutarch's style sometimes does shift to reflect that of his source,[2] then we might occasionally expect a text like that on Sertorius to shift toward a more Latinate placement of the direct object – a feature that Odlin called "borrowing transfer" (Odlin, 1990, p. 96) and which has been documented by applied linguists in unrelated languages (see Fortescue, 1993; Odlin, 1990, pp. 98–99, 104–07).

This logcontrast cluster analysis has confirmed the conclusions reached from the cluster analysis of the crude data and has indeed added to them. We now can see that some Greek texts of Plutarch and Marcus Aurelius appear more Latinate than had been previously noticed. We shall consider these texts as special cases in the following section which will attempt to decide objectively which language groups texts belong to.

Discriminant Analysis

In order to decide whether the subjective conclusions obtained from the above analyses are valid, we shall now carry out a discriminant analysis. Standard linear discriminant analysis uses linear contrasts of the data and in a similar complement to logcontrast principal components analysis, we can use logcontrasts of the data in logcontrast discriminant analysis. We shall assume that the prior probabilities of a text being in the Greek or Latin groups are the same for each group, and that the misclassification costs are equal.

Discriminant analysis also assumes that the covariance matrices of the data sets are equal. In order to test this we shall use Box's M statistic (Box, 1949; see also Krzanowski & Marriott, 1994, p. 230). However, this test is said to be sensitive to departures from Normality in the data. Tests for marginal Normality in our data are rejected,[3] so we shall use a permutation-based version of the test.

The permutation version of the bias-corrected Box's M test gives $p = 0.302$ with a 99% confidence interval of (0.260,0.347). There is thus no evidence to reject the hypothesis of equality of the covariance matrices, so we can proceed to the discriminant analysis.

Aitchison (1986), page 177, shows that a composition should be allocated to, in this case the Latin group, if

2. It is important to note that our data demonstrates that this was not always the case: Plutarch's *Roman Questions* and *Life of Camillus* and *Life of Cicero*, which certainly had Roman sources, conform to Greek usage.

3. Using the Watson test statistic (see Aitchison (1986), page 144), the hypothesis of Normality of x_{Gj1} is rejected at the 2.5% level while similar hypotheses for x_{Lj1} and x_{Lj2} are rejected at the 1% level. The other variables do not give rise to significant deviances from Normality.

$$(\underline{\widehat{\mu_L}} - \underline{\widehat{\mu_G}})' \widehat{\Sigma_4^{-1}} y - \frac{1}{2} \underline{\widehat{\mu_L}}' \widehat{\Sigma_4^{-1}} \underline{\widehat{\mu_L}} + \frac{1}{2} \underline{\widehat{\mu_G}}' \widehat{\Sigma_4^{-1}} \underline{\widehat{\mu G}} > 0, \quad (19)$$

where

$$y_{ijk} = \log (x_{ijk} / x_{ij4}), \quad (20)$$

$k = 1, \ldots, 3$, Σ_4 is as defined in (12) and $\underline{\widehat{\mu_L}}$ and $\underline{\widehat{\mu_G}}$ are the mean vectors for the Latin and Greek data respectively, defined as:

$$\widehat{\mu_{Lk}} = N_L^{-1} \sum_{j=1}^{N_L} y_{Ljk} \quad (21)$$

where $k = 1, \ldots, 3$ and μ_G is defined similarly. The first term of (19) can be expressed as a log-contrast of the composition x. Standardised log-contrast discriminant scores can be obtained by dividing (19) by an estimate of its standard error,

$$s = \{(\underline{\widehat{\mu_L}} - \underline{\widehat{\mu_G}})' \widehat{\Sigma_4^{-1}} (\underline{\widehat{\mu_L}} - \underline{\widehat{\mu_G}})\}^{1/2}. \quad (22)$$

Some texts have been identified as having unusual word-orderings for their language. We shall hold these texts out of the discriminant analysis and use them as test data after the discriminant scores have been established. The remaining data will be used to establish these scores.

In order to establish how our discriminant analysis might perform on new data, rather than just the data in our study, we shall carry out m-fold cross-validation (see for example Breiman, 1984), with the data being split into m groups, one of which is held out each time. With $m = 3$, one Latin text and two Greek texts are misclassified, representing a success rate of 94.1% and with $m = 5$, only two Latin texts are misclassified, a success rate of 96.1%. It appears that this discriminant function will work well on unseen data.

For our data then, the standard deviation is $s = 3.118$ and the discriminant score, D_{ij} may be computed as:

$$D_{ij} = 0.550 \log x_{ij1} + 5.281 \log x_{ij2} - 2.978 \log x_{ij3} - 2.852 \log x_{ij4} - 5.937. \quad (23)$$

We can now proceed to examine the results when the discriminant function (23) is applied to our held-out texts. The log-contrast principal components analysis above highlighted the texts of Cassius Dio and the Greek translation of *Res Gestae* as appearing to have a Latinate word order. The discriminant analysis produces positive scores for all of these texts, indicating that they have a Latinate word order. In the same analysis, Varro's *Res Rusticae* and Tacitus' *Agricola* appeared more Graecian. It can be seen from Table I that Varro's text is confirmed by the analysis as having a Latinate word order, while Tacitus' *Agricola* is assigned a Greek ordering.

The section on cluster analysis above also highlighted the Latinate nature of Plutarch's *Life of Sertorius* and Marcus Aurelius *Meditations* Book 3. These texts are also assigned to the Latin word-order group by the discriminant analysis.

Table 1. Held-out Texts in Increasing Order of their Discriminant Scores.

Author	Text	Key	Score	Conclusion
Tacitus	*Agricola*	L. 31	–1.197	Greek
Plutarch	*Life of Sertorius*	G. 18	0.145	Latin
Cassius Dio	Book 36	G. 2	0.276	Latin
Augustus	*Res Gestae*	G. 1	0.350	Latin
Marcus Aurielius	*Meditations* Book 3	G. 9	0.379	Latin
Cassius Dio	Book 54	G. 4	0.470	Latin
Cassius Dio	Book 59	G. 5	0.816	Latin
Varro	*Res Rusticae*	L. 34	0.849	Latin
Cassius Dio	Book 40	G. 3	2.318	Latin

CONCLUSIONS

We have shown that compositional data analysis is a better tool for analysing this kind of word-order data. The crude cluster analysis carried out by Frischer et al (forthcoming) picked out the texts of Cassius Dio and the translation of Augustus' *Res Gestae* as having Latinate word-order, despite being written in Greek.

Our new analysis confirms and extends these previous conclusions. With a more suitable method, we have also shown that Tacitus' *Agricola* has a Graecian word-order, possibly a reflection of its genre as the other texts by Tacitus in our sample are histories, while this is a biography. In addition, Book 3 of Marcus Aurelius' *Meditations* and Plutarch's *Life of Sertorius* appear to have Latinate word-orders despite being written in Greek. We noted in the section on cluster analysis that Marcus Aurelius was a native speaker of Latin and that Book 3 is known to be less revised and polished than the other books of the *Meditations*. Plutarch's sources for the *Life of Sertorius* were exclusively Latin and the text may reflect this.

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika, 70,* 57–65.

Aitchison, J. (1986). *The statistical analysis of compositional data.* Monographs in Statistics and Applied Probability. London: Chapman and Hall.

Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36,* 317–346.

Breiman, L., Friedman, J. H., Olsen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey: Wadsworth and Brooks.

Brunt, P.A. (1974). Marcus Aurelius in his Meditations. *Journal of Roman Studies, 64,* 1–20.

Chayes, F., Kruskal, W. (1966). An approximate statistical test for correlations between proportions. *Journal of Geology, 74,* 692–702.

Christ, W. (1905). *Geschichte der griechischen Literatur bis auf die Zeit Justinians* (4th edn.). Munich.

Corder, S. P. (1983). A role for the mother tongue. In S.M. Gass, and L. Selinker (Eds.). *Language transfer in language learning* (pp. 85–97). Massachusetts: Rowley.

Fortescue, M. (1993). Eskimo word order variation and its contact-induced perturbation. *Journal of Linguistics, 29,* 267–289.

Frischer, B.D., Andersen, R., Burnstein, S., Crawford, J., Dik, H., Gallucci, R., Gowing, A., Guthrie, D., Haslam, M., Holmes, D.I., Rudich, V., Sherk, R.K., Taylor, A., Tweedie, F.J., & Vine, B. (forthcoming). Word-order transference between Latin and Greek: The relative position of the accusative direct object and the governing verb in Cassius Dio and other Greek and Roman prose authors. Forthcoming in *Harvard Studies in Classical Philology*

Kellerman, E. (1983). Now you see it, now you don't. In Gass, S.M. & Selinker, L. (Eds.), *Language transfer in language learning* (pp. 112–134). Massachusetts: Rowley.

Krzanowski, W.J. & Marriott, F.H.C. (1994). *Multivariate analysis part 1: Distributions, ordination and inference,* (volume 2 of *Kendall's Library of Statistics*). London: Edward Arnold.

le Maitre, R.W.L. (1968). Chemical variation within and between volcanic rock series – a statistical approach. *Journal of Petrology, 9,* 220–252.

Norden, E. (1958). *Die antike Kunstprosa,* volume 1. Darmstadt, 5th edition.

Odlin, T. (1990). Word-order transfer, metalinguistic awareness, and constraints on foreign language learning. In B. van Patten, & J.F. Lee (Eds.), *Second language acquisition/ foreign language learning.* Phildelphia: Clevedon

Pelling, C.B.R. (1988). *Plutarch, Life of Anthony.* Cambridge: CUP.

Selinker, L. (1969). Language transfer. *General Linguistics, 9,* 67–92.

APPENDIX

Latin Data Set

Author	Text	Key	Direct Object position			
			MCB	OB	MCA	OA
Ammianus Marcellinus	Book 18	L.1	39.00	43.00	6.00	12.00
Ammianus Marcellinus	Book 22	L.2	41.00	41.00	12.00	6.00
Augustus	*Res Gestae*	L.3	63.00	27.00	6.00	4.00
Aulus Gellius	*Noctes Atticae* (Greek source)	L.4	35.00	52.00	8.00	5.00
Aulus Gellius	*Noctes Atticae* (Latin source)	L.5	37.00	49.00	10.00	4.00
Aulus Hirtius	*Bellum Gallicum*	L.6	37.00	45.00	7.00	11.00
Caesar	*Bellum Gallicum*, Book 1	L.7	48.00	48.00	3.00	1.00
Caesar	*Bellum Gallicum*, Book 3	L.8	46.00	51.00	3.00	0.00
Cato	*De Agricultura*	L.9	58.00	38.00	3.00	1.00
Celsus	*De Medicina*, Book 2	L.10	54.00	33.00	12.00	1.00
Cicero	*Ad Atticum*	L.11	35.00	46.00	10.00	9.00
Cicero	*Brutus*	L.12	24.00	57.00	6.00	13.00
Cicero	*De Legibus*	L.13	27.00	49.00	16.00	8.00
Cicero	*De Natura Deorum*	L.14	26.00	60.00	13.00	1.00
Cicero	*De Officiis*, Book 1a	L.15	27.00	53.00	9.00	11.00
Cicero	*De Officiis*, Book 1b	L.16	26.00	55.00	6.00	13.00
Cicero	*De Officiis*, Book 2a	L.17	24.00	53.00	13.00	10.00
Cicero	*De Officiis*, Book 2b	L.18	30.00	53.00	9.00	8.00
Cicero	*De Officiis*, Book 3a	L.19	23.00	56.00	11.00	10.00
Cicero	*De Officiis*, Book 3b	L.20	30.00	55.00	9.00	6.00
Cicero	*In Catilinam*	L.21	35.00	40.00	19.00	6.00
Livy	*Historiae*, Book 3	L.22	49.44	30.34	14.61	5.62
Livy	*Historiae*, Book 21	L.23	48.45	29.90	16.50	5.16
Pliny	*Naturalis Historia*, Book 34	L.24	31.58	41.05	18.95	8.42
Pliny	*Naturalis Historia*, Book 36	L.25	40.82	35.71	12.25	11.22
Pliny (Junior)	*Epistulae*	L.26	40.00	34.00	20.00	6.00
Seneca	*De Clementia*	L.27	38.14	42.27	14.43	5.16
Seneca	*Epistulae*	L.28	36.00	45.00	11.00	8.00
Scriptores Historia Augusta	*Life of Hadrian*	L.29	38.00	45.00	4.00	13.00
Scriptores Historia Augusta	*Life of Marcus Aurelius*	L.30	42.00	41.00	12.00	5.00
Tacitus	*Agricola*	L.31	65.31	13.27	9.18	12.25
Tacitus	*Annales*	L.32	47.00	42.00	5.00	6.00
Tacitus	*Historiae*	L.33	53.00	40.00	6.00	1.00
Varro	*Res Rusticae*	L.34	19.00	55.00	7.00	19.00

Note. For the passages and editions used see Frischer et al. (Forthcoming).

Greek Data Set

Author	Text	Key	Direct Object position			
			MCB	OB	MCA	OA
Augustus	*Res Gestae*	G.1	65.00	21.00	7.00	7.00
Cassius Dio	Book 36	G.2	55.00	27.00	9.00	9.00
Cassius Dio	Book 40	G.3	60.00	25.00	15.00	0.00
Cassius Dio	Book 54	G.4	75.49	14.71	5.88	3.92
Cassius Dio	Book 59	G.5	74.23	15.46	8.25	2.06
Dionysius Of Halicarnassus	*Roman Antiquities*, Book 2	G.6	37.62	22.77	14.85	24.75
Herodotus	Book 1	G.7	20.00	25.00	25.00	30.00
Herodotus	Book 2	G.8	20.00	35.00	21.00	24.00
Marcus Aurelius	Book 3	G.9	28.00	40.00	5.00	27.00
Marcus Aurelius	Book 8	G.10	28.00	31.00	21.00	20.00
Musonius Rufus	*Duscourses*, 1, 2	G.11	23.71	23.71	21.65	30.93
Plutarch	*Dinner*	G.12	14.00	44.00	12.00	30.00
Plutarch	*Greek Questions*	G.13	50.00	12.00	30.00	8.00
Plutarch	*Life Of Alcibibiades*	G.14	25.00	24.00	18.00	33.00
Plutarch	*Life Of Alexander The Great*	G.15	15.00	40.00	14.00	31.00
Plutarch	*Life Of Camillus*	G.16	40.00	21.00	28.00	11.00
Plutarch	*Life Of Cicero*	G.17	22.00	37.00	12.00	29.00
Plutarch	*Life Of Sertorius*	G.18	23.00	43.00	7.00	27.00
Plutarch	*Life Of Themistocles*	G.19	29.00	29.00	21.00	21.00
Plutarch	*Roman Questions*	G.20	33.00	25.00	22.00	20.00
Polybius	Book 1	G.21	21.00	26.00	17.00	36.00
Polybius	Book 2	G.22	20.00	24.00	9.00	47.00
Thucydides	Book 2	G.23	25.00	35.00	24.00	16.00
Thucydides	Book 5	G.24	40.00	13.00	36.00	11.00
Thucydides	Book 7	G.25	37.86	17.48	37.86	6.80
Xenophon	*Anabasis*	G.26	17.00	20.00	25.00	38.00

Note. For the passages and editions used see Frischer et al. (Forthcoming).