

# New Research on the Stylometry of Latin prose

## SESSION

*ORGANIZER: Bernard Frischer*

---

*Dept. of Classics, UCLA*

### AUTHORS:

Roger Andersen (UCLA)  
Jane Crawford (Loyola Marymount University)  
Bernard Frischer (UCLA)  
Ralph Gallucci (University of Central Arkansas)  
Donald Guthrie (UCLA)  
Ann Taylor (University of Pennsylvania)  
Emily Tse (University of Pennsylvania)  
Fiona Tweedie (University of the West of England)

### Organizer's Statement

As the 1995 conference of the ALLC-ACH in Santa Barbara showed, there is growing interest in exploiting the recently digitized corpus of Latin texts, particularly by scholars of style and authorship. At the 1996 conference we propose a three-paper session at which representatives of research teams will report on exciting new discoveries and work in progress in different sub-fields. Besides the focus on Latin prose, the three papers share a common methodology, which is based on statistical and stylometrical analysis.

(1) Bernard Frischer will report his team's discovery of a new distinguishing feature of Latin vs. Greek prose style. Cluster analysis reveals that whereas Greek prose authors place the direct object before or after the verb with equal frequency, Roman authors have a preference for placement before the verb. Frischer suggests that a current theory from Applied Linguistics can account for the fact that when Roman authors wrote in Greek, they often retained their native word order, and vice versa with Greek authors writing in Latin. The theory can also explain the related (and unexpected) observation that when Romans wrote in Latin utilizing a Greek source (and vice versa), they tended to shift their word order to reflect that of their source. For example, as cluster analysis proves, when the Greek writer Plutarch wrote biographies of Romans based on Latin sources, his Greek word order resembles the Latin pattern. But when he wrote about Greeks, his word order reverted to its normal Greek pattern. Frischer concludes by applying these insights to some problems of the native language of several ancient authors (including Plotinus and Charisius); and to the much-debated question of whether or not Book

3 of Cicero's *De Officiis* had a Greek source.

(2) Emily Tse then will report on her project to determine the validity of Dessau's thesis (first proposed in 1889) that the important late-antique collection of imperial biographies known as the *Historia Augusta* was written--not by the six authors mentioned in the manuscripts--but by a single author. She presents a summary and critique of recent work by Meissner, which attacked the Dessau thesis (made popular, after decades of controversy, in a pioneering computer study by Marriott in 1979) by focusing on rates of variance in single vs. multi-author corpora. Tse shows that although Meissner's study is flawed, it may well have put us on the right track to settling the question once and for all, while also providing a useful tool (variance of function words) for helping us to distinguish single from multiple-author corpora in Latin and, very possibly, in other languages as well.

(3) Fiona Tweedie then will take us forward a millennium to the Neolatin text known as the "De Doctrina Christiana", which some scholars have attributed to Milton. Debate about this question is fierce. Tweedie's work promises to make an important contribution. Working with digitized versions of this text and others by Milton and his contemporaries, Tweedie and her team identify discriminators that can separate the known Milton samples from the control texts to see where "De Doctrina Christiana" falls. In this paper, Tweedie will concentrate on the frequency of function words. First, Tweedie shows by a principal components analysis that study of the one hundred most frequently occurring words does successfully discriminate the known Miltonic and control texts. Then she studies where the "De Doctrina" falls among Milton's works. The results on each of her various tests differ, but in all cases the "De Doctrina" is seen to overlap with other works of Milton. The conclusion reached is that the author of the "De Doctrina Christiana" may be Milton, but literary genre is also a strong factor.

# Word-Order Transference Between Latin and Greek

*Bernard Frischer, Roger Andersen<sup>1</sup>,  
Jane Crawford<sup>2</sup>, Ralph Gallucci<sup>3</sup>, Do-  
nald Guthrie<sup>4</sup>, Emily Tse<sup>5</sup>, Ann Taylor<sup>6</sup>*

<sup>1</sup>*Dept. of Classics, UCLA, 405 Hilgard Avenue,  
Los Angeles, California, USA 90095-1475*

<sup>2</sup>*Dept. of Classics, Loyola Marymount Universi-  
ty, Loyola Blvd.at W. 80th, Los Angeles, CA  
90045-2699*

<sup>3</sup>*University of Central Arkansas*

<sup>4</sup>*Dept. of Biostatistics, School of Public Health,  
UCLA, 405 Hilgard Avenue, Los Angeles, CA  
90095-1759*

<sup>5</sup>*Dept. of Classics, University of Pennsylvania,  
1023 S. Hudson Avenue, Los Angeles, CA 90019*

<sup>6</sup>*Dept. of Linguistics, University of Pennsylvania,  
Williams Hall, Philadelphia, PA 19104*

KEYWORDS: Word order, transference, transfer to  
somewhere theory (TTS)

AFFILIATION: <sup>1</sup> Dept. of Classics, UCLA

<sup>2</sup> Dept. of Classics, Loyola Marymount University

<sup>3</sup> University of Central Arkansas

<sup>4</sup> Dept. of Biostatistics, School of Public Health,  
UCLA

<sup>5</sup> Dept. of Classics, University of Pennsylvania

<sup>6</sup> Dept. of Linguistics, University of Pennsylvania

E-MAIL: iddhbdf@mvs.oac.ucla.edu

FAX NUMBER: (310) 206-2471

PHONE NUMBER: (310) 391-4068

## Word-order Transference between Latin and Greek

Our research focuses on a feature of Greek and Latin style that was not readily apparent before the corpus of ancient texts was digitized and analyzable by various text processing programs.

Applied linguists have observed that learners of second language often transfer features of their first language to the language they are studying (cf. S. Gass and L. Selinker 1983, Andersen and Shirai 1994). These features may range from the phonological to the morphological or syntactic. In applied linguistics, the reason for studying such effects (sometimes called “input bias”) is pragmatic: by understanding the kinds of errors that typically arise in language learning, it is hoped to streamline the process of language acquisition.

In this paper, we study one example of such transference for Greek and Latin concerning an aspect of word order in Greek and Latin: the placement of the direct object with respect to the main verb. First we establish a significant difference in the

Greek and Latin distributions. Whereas Greek writers tend to place the direct object before and after the main verb with more or less equal frequency, Roman writers have a distinct tendency to place the direct object before the main verb. Evidence for this is presented from the following writers: for Greek, Diodorus Siculus, Diogenes Laertius, Eunapius, Eusebius, Herodotus, Philostratus, Polybius, Thucydides, Xenophon; for Latin, Caesar, Cicero, Historia Augusta, Livy, Tacitus. Searches of relevant recent bibliographies (Janse, 1994; Werner, 1994) and personal communications from Classical linguists suggest that this difference in the Greek and Latin distributions has not yet been noted. Preliminary data suggest that the patterns found for the direct object also hold for the indirect object.

Next we show that when Greek authors write in Latin, they usually preserve their native Greek word-order pattern; and this is also seen to happen when Latin authors write in Greek. For Greek, we use the examples of the Roman historian Dio Cassius, the emperor-philosopher Marcus Aurelius, and Clemens, bishop of Rome; for Latin, the example of the Greek historian, Ammianus Marcellinus.

Our theory to account for these observations is as follows. Since both Latin and Greek had relatively free word order, there was no definite “right” and “wrong” position of the direct object with respect to the main verb. In each language, the direct object can either precede or follow the main verb. Thus, the second of Andersen’s two conditions for the transfer to somewhere principle (TTS) is fulfilled: “(1) natural acquisitional principles are consistent with the native language structure or (2) there already exists within the second language input the potential for (mis-)generalization from the input to produce the same form or structure” (Andersen 1983; Andersen 1990). In learning the other language, Latin and Greek speakers were probably not corrected for their DO placement in any individual utterance. That the overall distributions in the two languages were quite different is not something that was ever consciously observed in antiquity; nor was it readily apparent before the application of computing to Classical philology through the digitization of the entire corpora of Latin and Greek and the application of statistical analysis to stylistic features readily retrieved from the corpora. Indeed, the ancient grammarians stressed the congruence and even consanguinity of the two languages, sometimes viewing Latin as derived from Greek or even as a form of the Aeolic dialect (cf. R. Giomini 1953; E. Gabba, 1963; K. Schoepdsau 1992).

Having seen that the author’s native word-order pattern biases his style when he writes in the other language, we turn to a related, but unexpected,

analogous case of input-bias: when a writer of one of the languages writes in his native language using a source in the other language, his word-order preferences begin to shift to the word order typical of the source language. This kind of input-bias we call "cross-influence." The most extreme cases occur in translations (see, e.g., A. Debrunner and A. Scherer 1969, for Latinisms in Greek translations generally); but it is also often operative when an author writing in one language simply uses a source from the other. Examples discussed for translation include: Aulus Gellius (cf. P. Steinmetz 1992), Cicero (cf. C. Mueller-Goldingen 1992), and the *Res Gestae Divi Augusti*. For the input-bias of sources we discuss: the Greek and Roman lives of Cornelius Nepos (a Latin writer) and Plutarch (a Greek writer). Cluster analysis is used to show that the Roman biographies of Plutarch (though written in Greek) have distributions resembling the Roman lives of Nepos; whereas the Greek lives of Nepos (though written in Latin) have distributions approximating those of the Greek biographies of Plutarch. Our claim is purely empirical: not that cross-influence *must* occur whenever an author uses a source in the other language, but simply that such input-bias can be documented in the cases we have studied.

Our theory for explaining cross-influence again is based on Andersen's theory of input-bias. In this case, however, the input comes not from the writer's own native *Sprachgefuehl* but from the syntax of his source, which evidently exerts an unconscious, attractive force on the author of the derivative text, who imitates not only the content but also the style of the source-text. (An interesting analogous case of the subtle influence of English word order on Eskimo is discussed by M. Fortescue, 1993.)

Once these transference effects have been established, we suggest that they can be utilized--not for instructional purposes, as has been the case with Applied Linguistics and living languages--but for the solution of some philological problems. For example, we study the distributions of Books I and II of Cicero's *De Officiis* (which we definitely know had a Greek source) and compare them to what is found in Book III (where the question of a Greek source is controversial; see Dyck 1996). We use a similar approach to determine the likelihood that several ancient authors of unknown origin were Greeks or Romans (e.g., Greeks (?) writing in Latin: Euanthius and Charisius; see P. L. Schmidt 1989; Roman/Greek (?) writing in Greek: Plotinus).

## References

- Andersen, R., 1983: "Transfer to Somewhere," *Language Transfer in Language Learning*, edited by Susan Gass & Larry Selinker, Boston: Newbury House Publishers, pp. 177-201.
- Andersen, R., 1990: "Models, Processes, Principles & Strategies: Second Language Acquisition Inside and Outside the Classroom," in *Second Language Acquisition/Foreign Language Learning*, ed. B. VanPatten and J. F. Lee, Clevedon and Philadelphia, pp. 45-68.
- Andersen, R. and Shirai, H., 1994: "Discourse Motivations for Some Cognitive Acquisition Principles," *Studies in Second Language Acquisition* 16.2, pp. 133-156.
- Debrunner, A. and Scherer, A., 1969: *Geschichte der griechischen Sprache*, vol. 2, Berlin.
- Dyck, A., 1996: *Cicero's De Officiis. A Commentary*, Ann Arbor.
- Fortescue, M., 1993: "Eskimo Word Order Variation and Its Contact-Induced Perturbation," *Journal of Linguistics* 29, pp. 267-289.
- Gabba, E., 1963: "Il latino come dialetto greco," in *Miscellanea di studi alessandrini in memoria di Augusto Rostagni*, Turin, pp. 188-194.
- Gass, S. and Selinker, L., eds., 1983: *Language Transfer in Language Learning*, Rowley, Mass., London, Tokyo.
- Giomini, R., 1953: "Il grammatico Filosseno e la derivazione del latino dall'eolico," *La Parola del Passato* 8, pp. 365-376.
- Janse, M., 1994: "L'ordre des mots dans les langues classiques. Bibliographie des années 1939-1993," *Tema. Techniques et Methodologies modernes appliquées a' l'Antiquité* 1, pp. 187-211.
- Mueller-Goldingen, C. 1992: "Cicero als Uebersetzer Platons," *Zum Umgang mit fremden Sprachen in der griechisch-roemischen Antike, Palingenesia* 36, 173-188.
- Schmidt, P. L., 1989: "Grammatik und Rhetorik," in *Restauration und Erneuerung. Die lateinische Literatur von 284 bis 374 n. Chr.*, ed. R. Herzog, Munich, pp. 101-214.
- Schoepsdau, K., 1992: "Vergleiche zwischen Lateinisch und Griechisch in der antiken Sprachwissenschaft," in *Zum Umgang mit fremden Sprachen in der griechisch-roemischen Antike, Palingenesia* 36, pp. 115-136.
- Steinmetz, P., 1992: "Gellius als Uebersetzer," *Zum Umgang mit fremden Sprachen in der griechisch-roemischen Antike, Palingenesia* 36, pp. 201-212.
- Werner, J., 1992: "Bibliographie zur Problematik der Fremdsprachlichkeit in der griechisch-roemischen Antike," *Zum Umgang mit fremden Sprachen in der griechisch-roemischen Antike, Palingenesia* 36, pp. 233-252.

# Is Variance of Function Words a Reliable Discriminator of Single and Multiple Author Corpora of Latin Prose? An Empirical Critique of Meissner's Studies of the *Historia Augusta*.

*Emily Tse, Bernard Frischer*

---

Box 421 HRN, 3901 Locust Walk, Philadelphia, PA 19104-6135

KEYWORDS: variance, authorship studies

AFFILIATION: Emily Tse, University of Pennsylvania; Bernard Frischer, UCLA

E-MAIL: Emily Tse, [etse@mail.sas.upenn.edu](mailto:etse@mail.sas.upenn.edu)

FAX NUMBER: 215.573.7794  
(c/o Classics Department)

PHONE NUMBER: 215.417.8858

The problem of the *Historia Augusta* has been the subject of much debate for many decades. This biographical collection of Roman emperors covers a period from AD 117-285 and is attributed in the mss. to six different authors. In 1889 however, Dessau, having studied the nomenclature and style of the works in the HA, proposed the theory that it was written only by one author.

Almost a century later, Marriott offered support for Dessau's theory through stylometric analysis. His first study compared the average number of words per 'sentence' of the *Historia Augusta* with those of other fourth century texts such as the *Codex Theodosianus*. His second study made a similar comparison based on the choice of word-type (part of speech) at the beginning and end of sentences. Both Marriott's studies showed similarities among the biographies of the HA but differences from the control texts. Based on these findings he concluded that the collection was authored by one person, as Dessau had proposed.

Recently, Frischer, et al. published critiques of Marriott's work by extending the study to include control texts which fall under the same genres of the *Historia Augusta*, namely biography and history. Results that are worth noting surfaced: (1) The average numbers of words per 'sentence' in the HA and the control texts were fairly close. Marriott's averages, which included generically-unrelated control texts, had a broader range. (2) Tests by word-type also showed stylistic similarities

between the HA and such authors as Livy, Tacitus, and Suetonius. On the other hand, Marriott's two control texts for his second study turned out to be eccentric. These results tend to question the validity of Dessau's theory of single authorship as well as Marriott's methods for stylometric investigation.

Thus, to make further progress in assessing Dessau's thesis, a test is needed which is sensitive enough to detect stylistic features specific to the author and not merely the tradition in which he and his colleagues wrote. One potentially powerful tool is function word analysis, on which Meissner published a paper in 1992.

In his study, Meissner compared the frequency of certain function words in the *Historia Augusta* to those found in Suetonius' *De Vita Caesarum*. He specifically focused on those words which appeared most frequently throughout the texts. These function words are the following seven: ad, cum, est, et, in, non, and ut. Meissner then ran statistical tests on the results to address two concerns. He first questioned whether the fluctuating behavior of the frequencies in the HA could be interpreted as random or pattern-like. As a result, Meissner ran chi-square tests. The values calculated, by Meissner's reasoning, suggest that the text is homogeneous and from one source.

Meissner's second concern was quantifying the direct relationship between the frequencies found in the HA and those in Suetonius. The former should fluctuate in a similar way to the latter, if the *Historia Augusta* was indeed written by a single author. To address this issue, Meissner studied their variances and ran F-tests which are the ratios between the variances of the two samples. The two samples in our case, are texts of the HA and Suetonius. If they are from the same population, their variances should be similar. For our purposes, we regard the same population as two texts both being drawn from single-authored corpora or both from multiple-authored corpora. If F is a much greater value, however, then no substantial likeness between the two texts has been detected. So this F-test provides a numerical standard by which we can consider whether two texts come from the same source.

In Meissner's study, the variances for each of the function words in the *Historia Augusta* were larger than Suetonius', exhibiting a larger fluctuation and dispersion. Similarly, the F-values were quite large, allowing us to reject a similar source for the two texts. Thus, Meissner concluded the HA is not single-authored, as the *De Vita Caesarum* is.

Meissner's use of function word analysis seems sensitive and subtle enough to detect differences between the HA and Suetonius; and all the statistics appear to point to a multi-authored *Historia Augusta*. However, the study is not empirical

enough, since it used only Suetonius as a basis for comparison. Some data were gathered on Nepos, but were not fully exploited.

The new study presented here strives to offer more empirical evidence by investigating the frequencies of the function words in the works of the historians Livy and Tacitus. We begin by noting the frequencies of each of the seven function words per text. From each of the four extant decades of Livy's *Ab Urbe Condita*, five books were taken. Similarly, for our sample from Tacitus, in addition to his three non-historical pieces, we take five books from the *Annales* and five from the *Historiae*.

Next we test for randomness, as Meissner did, by conducting chi-square tests on the data, along with the revised ones on the HA and Suetonius. We also examine their p-values. The probability for Suetonius is .118 while the HA's is extremely low, which rejects the null hypothesis for independence and seems to confirm Meissner's studies. On the other hand, when we observe the values for Livy and Tacitus, we find results that do not confirm Meissner. Both p-values are .000. These are values we do not expect, since the texts were both written by one author.

In Meissner's study, the relationship of the frequencies between texts was also quantified by calculating their variances and running the F-test. In order to better study the relationship between single-authored corpora and the HA, multi-authored corpora were artificially created in the study. By including these newly created works for the F-tests, we can judge how reliable a discriminator the variance of function words really is. A number of results do not correspond to our expectations. The p-values from chi-square tests on Livy and Tacitus both come out to be .000, which, according to Meissner's reasoning, might suggest that some of the texts in their corpora were not written by these two authors. One explanation may be that the values are reflecting a change in style over time--a possibility that is not improbable when we examine their frequencies by percent.

The percentages also indicate that the frequencies remain fairly constant throughout Tacitus' *Historiae* and *Annales*. However, we detect some fluctuation patterns in the books of Livy for certain function words. These fluctuations can be a cause for the unexpected x2 values found. Whatever the case may be, the problem with the chi-square test is that it is too sensitive. By observing differences within a homogeneous text, it makes distinctions in more ways than we had intended. For this reason, the chi-square can be misleading and should not have been included in Meissner's study. Although the F-test does not make these kinds of fine distinctions, we encounter difficulty with the values, which are generated at a 29% rate of error.

One problem is that the null hypothesis is ambiguous, stating that two samples come from the same population. The same population, for Meissner's purposes, is single-authored or multi-authored corpora. However, it can also undeniably be interpreted as one author vs. another, such as Livy vs. Tacitus. So, in fact we have two possible hypotheses operating here.

Another possibility for re-interpretation is to look at specific function words that may reflect what we expect, rather than at the mean F-ratios. When we reexamine the values, we find that those for *est* and *ut* are excellent examples. The F-ratios are lower in known single vs. known single author corpora as well as in the HA vs. another known multiple author work. Similarly, the ratios are higher in known single vs. known multi-authored texts. It is likely that *est* and *ut* may serve as "magic keywords" in the analysis of function words. Because of the structure of the Latin language, *est* and *ut* are syntactically multi-functional. The remaining five words, however, are much less so. Thus, it is not surprising that *est* and *ut* are better discriminators in tests of variance. In conclusion, Meissner's study of variance of function words is flawed but promising. Some issues in the theory behind it still remain unresolved, such as the effects of composition over a long period of time. Yet, it would still be worthwhile to continue exploring the reliability of variance of function words. For instance, the emergence of *est* and *ut* as powerful tools needs further investigation. In addition, there is the possibility of other 'multi-functional words' which may help us detect single and multiple authorship. Given these avenues for further analysis, variance of function words may still prove itself a powerful discriminator.

## References

- Agresti, A. and B. Finlay (1986), *Statistical Methods for the Social Sciences* (San Francisco).
- Bird, H. W. (1994), *Aurelius Victor: De Caesaribus. Translated with an Introduction and Commentary* (Liverpool).
- Dessau, H. (1889), "Ueber die Zeit und Persoenlichkeit der SHA," *Hermes* 24:337-392.
- Dessau, H. (1892), "Ueber die SHA," *Hermes* 27: 561-605.
- Frischer, B. (1991), *Shifting Paradigms. New Approaches to Horace's Ars Poetica* (Atlanta).
- Frischer, B. (1995), "How to Do Things with Words per Strong Stop: Two Studies on the Historia Augusta and Cicero," in *Aspects of Latin. Papers from the Seventh International Colloquium on Latin Linguistics, Jerusalem 19-23 April 1993*, ed. by H. Rosn. Forthcoming in *Innsbrucker Beitrge zur Sprachwissenschaft* 1995.

- Frischer, B., et al. (1996), "'Sentence' Length and Word-type at 'Sentence' Beginning and End: Reliable Authorship Discriminators for Latin Prose? New Studies on the Authorship of the *Historia Augusta*," forthcoming in *Research in Humanities Computing*, vol. 6.
- Horsfall, N. (1989), *Cornelius Nepos: A selection, including the lives of Cato and Atticus. Translated with an Introduction and Commentary* (Oxford).
- Lindsay, K. L. and Mackay, T. W., "An Authorship Study of the Pauline Epistles," an unpublished paper given at the International Conference on Computers in the Humanities (Brigham Young University, June 26, 1985) 1-33.
- Marriott, I. (1979), "The Authorship of the *Historia Augusta*: Two Computer Studies," *JRS* 69:65-77.
- Meissner, B., "Computergestützte Untersuchungen zur stilischen Einheitlichkeit der *Historia Augusta*," forthcoming in *BHAC*.
- Meissner, B., "Sum enim unus ex curiosis. Computerstudien zum Stil der *Scriptores Historiae Augustae*," *RCCM* 34: 47-79.
- Panhuis, D. (1984), "Is Latin an SOV Language?" *Indogermanische Forschungen* 89:140-159.
- Schenkeveld, D.M. (1964), *Studies in Demetrius On Style* (Amsterdam).

## The Provenance of Christian Doctrine, attributed to John Milton: An Evaluation of Alternative Statistical Methods

*F. J. Tweedie, T. N. Corns, J. K. Hale, G. Campbell and D. I. Holmes*

*Department of Mathematical Sciences, Frenchay Campus, Coldharbour Lane, BRISTOL, BS16 1QY, UK*

AFFILIATION: University of the West of England, Bristol, UK; University of Wales, Bangor, UK; University of Otago, New Zealand; University of Leicester, UK; University of the West of England, Bristol, UK

E-MAIL: fiona.tweedie@uwe.ac.uk  
 FAX NUMBER: +44 117 976 3860  
 PHONE NUMBER: +44 117 965 6261 Ext 3175

### Introduction

This project attempts to resolve perhaps the most urgent issue in Milton Studies, namely the provenance of the Latin manuscript, found among the

English State papers some 150 years after Milton's death, which has since its discovery been regarded as his definitive attempt to define Christian Doctrine.

The inclusion of the text in the Milton canon has never been unproblematic. Initial responses to it regarded the text as a sensational disclosure of a heretical tendency almost unsuspected previously. In the mid and late twentieth century many significant studies have remarked on and attempted to explain away discrepancies between that text and other work most certainly by Milton, particularly "Paradise Lost" (for example Hunter et al. 1971). In recent years, William B. Hunter, who formerly had sought sophisticated explanations for those contradictions, became convinced that the text is that of an unknown seventeenth-century writer and that it was mistakenly or cynically foisted on Milton shortly after his death, an error compounded on its rediscovery (Hunter 1992, 1993). A group of scholars, Miltonists and statisticians, have been drawn to the question (Campbell et al. 1995), and the text is currently under interrogation from several perspectives, including a study of the amanuenses, a study of the possible circumstances of its composition and early transmission, a study (in part computer-aided) of its Latinity and a stylometric analysis.

From the earliest point in the history of stylometry, work has been performed on works in many different languages, including French, Swedish, Russian, Hebrew, and Greek. However, despite the interest shown in the latter two cases involving classical languages, Latin has remained relatively untouched by stylometric hand. Greek, on the other hand, has attracted much attention, perhaps by virtue of its New Testament biblical connections. Work has been carried out on the New Testament, poetry, and letters, amongst others. However, despite the fact that both Latin and Greek are inflected languages, there are fundamental differences. For example, the lack of the article in Latin invalidates the enthusiasm for sentence length investigations in Greek writings, see Michaelson et al (1978) and Wake (1957). In this paper, a development of the work presented as a poster at the ACH/ALLC conference in Santa Barbara in 1995, we investigate the application of various stylometric techniques. Before they can be applied to "De Doctrina Christiana" it is important to validate their effectiveness on texts of known authorship.

### "De Doctrina Christiana"

A group of scholars has formed with the intention of investigating the authorship of "De Doctrina Christiana". This neo-Latin manuscript was found in 1823 along with State Papers by Milton and it is this location that prompted the Miltonic attribu-

tion. The investigation has proceeded along various different lines, but the stylometric work is detailed in the next section.

### Stylometric work

Despite the lack of interest shown by the stylometric community, some research has taken place in combining statistics and Latin, see for example Hubka (1985). With a few exceptions, it remains amateurish and open to attack on both statistical and literary grounds. Much of it is detailed in Tweedie et al. (1995) We were able to obtain a machine readable version of Milton's Latin prose, in particular the three Defences, "Defensio Prima", "Defensio Secunda" and "Pro Se Defensio". These are the only comparable works by Milton in neo-Latin and fall into the polemic genre. We have also entered text samples from other texts for use as control samples. Three theological samples were chosen, by Ames, Wolleb and Baxter. The works by Ames and Wolleb were mentioned in "De Doctrina Christiana". Five polemic samples were also identified, works by May, Prynne, Wentworth, Earle and Bate. We also had access to samples of text by Bacon from the Oxford Text Archive. The Milton polemics, "Defensio Secunda" and "Pro Se Defensio", were of the order of 25,000 words and were split into five samples, while "Defensio Prima" has around 47,700 words and was split into nine samples for analysis. "De Doctrina Christiana" is very much larger and we only consider the first 25,000 words, again split into five samples. The control texts all have around 3000 to 5000 words and were kept as one sample each.

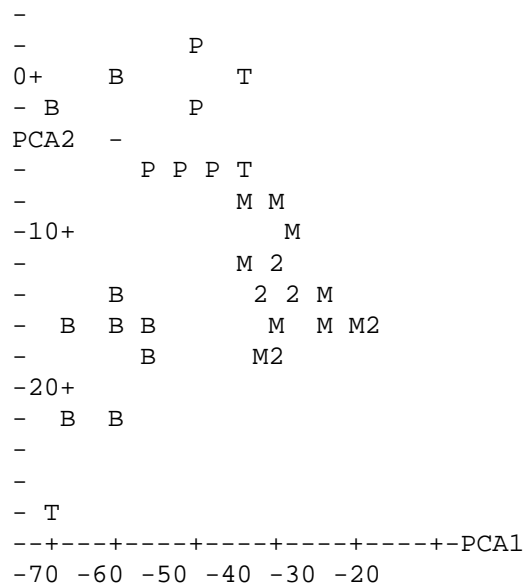
The aim of this section is to identify discriminators that are able to separate the known Milton samples from the control samples, and examine where "De Doctrina Christiana" falls on a similar scale. Measures employed so far include function word doubletons and the most common words used. The function word analysis is detailed in Tweedie et al (1995), here we concentrate on the analysis of common words.

### Most common words

In a technique similar to that of Burrows (1992), we identified the one hundred most frequently occurring words in the corpus of text. Their frequencies were standardized and then examined using principal components analysis. The analysis proceeded in two distinct branches. Firstly, it was important to discover if this technique was applicable to neo-Latin, by applying it to texts of known authorship. Secondly, we wished to investigate the relative discriminatory ability of sections of words, the top fifty, second fifty, top twenty-five and so on.

Initially, therefore, we considered only the texts of

known authorship, the Milton polemics, the polemic and theological control texts as well as the Bacon samples. The fifty most often occurring words were used as input to the principal components analysis which resulted in the graph below.

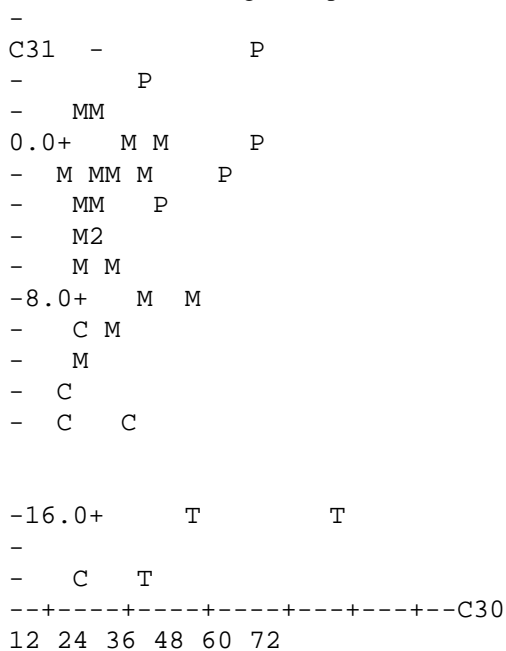


The samples are designated M for Milton polemics, B for Bacon samples, P for polemic controls and T for theological controls. It is clear that the Milton samples are closely clustered together in the right of the graph, while the Bacon samples are towards the left. With the exception of the sample from Baxter, the rest of the polemic and theological control samples are towards the centre of the plot. The clustering of the polemic and theological samples is especially interesting when it is remembered that they are all written by different authors. Indeed, it is clear that the first principal component separates Milton from the polemic/theological controls, as well as the Bacon samples. This component is based mainly on the use of 'et' ( $r=-0.98$ ). In Latin the enclitic 'que' is equivalent to 'et', but we have not counted -que usage. This appears to be a reason for the very low usages of 'et' evident in the Milton texts. The Bacon corpus, on the other hand, has a very low incidence of '-que'. Other words that are important on this axis include 'quid', 'quo', 'quidem' and 'qui', all words that Milton uses much more than Bacon. Stylistically, Bacon was known to write in the plain style, while Milton used more complex sentences. Our results would seem to confirm this.

The clustering of the Milton samples and their separation from the other controls indicates that this technique would be capable of discriminating between Milton and texts by other authors. We are therefore confident in applying it to the text of "De Doctrina Christiana". As mentioned above, the most commonly occurring one hundred words

were enumerated. For computational reasons, we were unable to analyse all one hundred words at once. It was therefore decided to split the data into, firstly two, the top fifty, then the next top fifty, and secondly into four, considering the top twenty-five words, the next twenty-five and so on. The data from De Doctrina was included in the analysis, represented by 'C's, and that from the Bacon samples was removed as it was not directly comparable for these purposes.

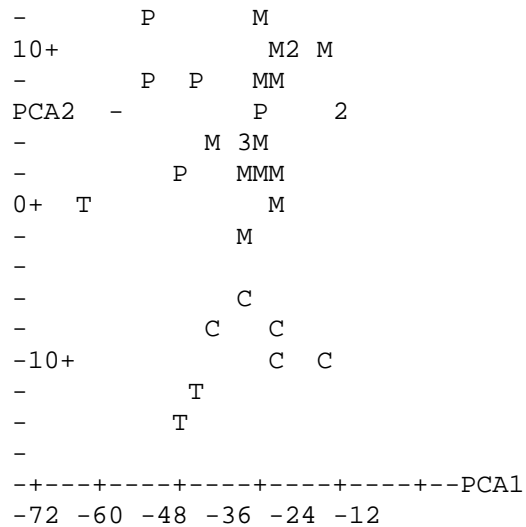
The first analyses were carried out on the twenty-five word samples. Analysis of the most common twenty-five words produced the graph below. It can clearly be seen that the Milton samples are in the top left area of the graph, the "De Doctrina Christiana" samples in the bottom left, and the controls towards the right, the polemics at the top.



The discriminations are not complete, there are "De Doctrina Christiana" samples mixed with the Milton, and a Milton sample towards the control area. However, the general structure is clear. Analyses of the other twenty-five word sections resulted in graphs that generally separated the groups, with various facets being demonstrated.

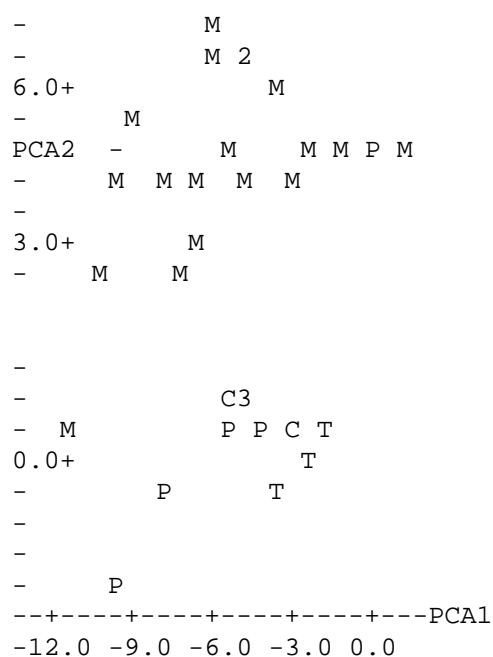
The second stage of the analysis was carried out using the top fifty words. The resulting plot is shown below. It appears that there are four distinct areas of the graph. The upper half contains the polemic samples, both controls and the Milton Defences, while the lower half contains the theological samples, "De Doctrina Christiana" and controls. The exception is again the Baxter sample, which appears higher than it perhaps should for its genre. Also, the right hand side of the graph appears to contain the Milton samples as well as "De Doctrina Christiana", while the left hand side seems to contain the control samples from both genres. This is a development of the above graph

from the top twenty-five words. Here the groups are separated better, and clustered more tightly together.



This PCA plot indicates that the "De Doctrina Christiana" samples are closer to Milton's defences than they are to the theological control samples. This would appear to lead to the conclusion that the author of "De Doctrina Christiana" uses these words in a similar way to that of Milton.

The analysis of the second fifty most common words results in the plot below. It can be seen that the Milton samples are spread over a large area of the first principal component, but in the higher part of the graph, while the other samples, with the exception of one of the polemic samples, Eikon Basilike, are in the lower section. The "De Doctrina Christiana" samples are closely grouped, as are the theological samples, in the lower, right corner of the graph.



Investigation of the words affecting the first principal component reveals that 'rex' plays the major part. Thus works of a polemic nature, with reference to the king, score more negatively than works of a theological nature. This may explain the close groupings of the theological samples. On the second principal component, analysis is more difficult. Many words play roles here, but the most significant are 'ego' and 'quis', as well as 'jam'. Again, the Milton samples appear different from the control samples and "De Doctrina Christiana", perhaps due to the complex sentence structure mentioned above.

### Conclusions and Future Work

It is clear that this area represents almost virgin territory for the stylometrician. Previous work has been reviewed and found wanting in many areas. The results of applying more appropriate techniques to the case of "De Doctrina Christiana" seem to imply that the author of "De Doctrina Christiana" may well be Milton, but genre effects appear to be playing a part. This was revealed by the analysis of the most common fifty words. Consideration was also given to the second top fifty and tranches of twenty-five words. Other techniques should be investigated, but further investigation is needed into Latin stylometry before more can be done.

Some aspects of further work would be unable to be carried out without access to a morphological parser, to identify parts of speech automatically, and a metrical scanner. This would enable the rhythm of Latin prose to be taken into account in investigations. Certain works have been scanned, but they are limited. Access to these utilities would dramatically open up the subject to further research.

### References

- Burrows, J. F. (1992). "Not Unless You Ask Nicely : The Interpretative Nexus between Analysis and Information." *Literary and Linguistic Computing*, 7(2):91--109.
- Campbell, G. C., Corns, T. N., Hale, J. K, Holmes, D. I. and Tweedie, F. J.. (1995). "The Provenance of 'De Doctrina Christiana': An Interim Report." International Milton Seminar, Bangor, UK, July 1995.
- Hubka, K. P., (1985). "Stylometric Test for Authorship of a supposed Comenianum, Explicatio Causae Moventis Naturalis." *Humanistica Lovaniensia*, 35:159--168.
- Hunter, W. B. (1992). "The Provenance of Christian Doctrine." *SEL*, 32:129--142.
- Hunter, W. B. (1993). "The Provenance of the Christian Doctrine: Addenda from the Bishop of Salisbury." *SEL*, 33:191--207.

- Hunter, W. B., Patrides, C. A. and Adamson, J. H. (1971). *Bright Essence: Studies in Milton's Theology*. University of Utah Press.
- Michaelson, S., Morton, A. Q. and Wake, W. C. (1978). "Sentence length in Homer and hexameter verse." *Association for Literary and Linguistic Computing Bulletin* 2.
- Tweedie, F. J., Corns, T. N., Hale, J. K., Campbell, G. and Holmes, D. I. (1995). "Latin and Stylometry: 'De Doctrina Christiana.'" *JADT*, Rome.
- Wake, W. C. (1957). "Sentence Length Distributions of Greek authors." *Journal of the Royal Statistical Society Series A* 120:331-346