Disputed Authorship: 30 Biographies and Six Reputed Authors.
A New Analysis by Full-Text Lemmatization of the 'Historia Augusta'

Dr. Penelope J. Gurney and Lyman W. Gurney

Dr. Penelope J. Gurney, Faculty of Education, University of Ottawa, 145 J. J. Lussier, Ottawa, Ontario, CANADA K1N 6N5
Lyman W. Gurney, Themis Research Corporation, Suite 507, 500 Laurier Ave West, Ottawa, Ontario, CANADA K1R 5E1

KEYWORDS: stylometry, authorship, lemmatization

AFFILIATION: Dr. Penelope J. Gurney, University of Ottawa; Mr. Lyman W. Gurney, Themis Research Corporation

E-MAIL: pgurney@uottawa.ca
FAX NUMBER (613) 562-5146
PHONE NUMBER (613) 562-5800 ext 4112

Introduction

This paper describes current research on authorship attribution that we are conducting by means of a computer assisted analysis of the thirty biographies in the 'Historia Augusta', a set of lives of Roman emperors and usurpers of the second and third centuries of this era.

The programming system that we have developed for research in textual analysis to provide these solutions has been used specifically, in this case, for an analysis of the original Latin, but has also been used successfully in analysis of English text: it is language independent. The system permits the production of a word-by-word comparison of the vocabularies of several texts, in which the actual words are presented either in original form or as lemmas, in order to permit immediate stylometric analyses of the given works. In this paper we demonstrate the results of some of the analyses which we have undertaken using its assistance.

Textual Analysis

Over the past twenty years, the uses of computer methods in the analysis of texts have grown both in scope and in numbers. The extensive overview of authorship attribution measures by Holmes (1994) discusses in detail methods which have been used in content analysis. It is clear from the discussion, however, that it may be possible to improve upon some of the techniques used in the past. For example, in many cases, a small set of function words, or a pre-selected small subset of commonly-occurring words, has been used to describe the attributes of an entire text. Given the power available in the standard office or home computer, however, unless there is compelling reason for the choice of a particular sample, it is no longer necessary to restrict one's efforts to small samples of a full vocabulary. Various subsets of available vocabulary will undoubtedly be used finally in any research project, but those subsets can be selected after full consideration of the entire vocabulary, and can be capable of immediate modification at any point in the research cycle.

The Historia Augusta

The problem which is inherent in the Historia Augusta lies in the necessity of attempting to reconcile the conflicting claims of the manuscript tradition, which names six authors (otherwise quite unknown) of the late third and early fourth centuries of this era, with the results of very intense literary and historical analyses of the past 100 years, which point quite strongly to an origin in the late fourth century, or even to the very early fifth century, and to perhaps but one author. This study will refer to 'authors', without being committed to any particular number.

The question of the authorship of this text must surely be one of the ultimate challenges in such studies, for the six authors occasionally mention each other, claim for themselves lives which have been attributed in the manuscripts to others of the six, in a few lives make most unlikely dedications to the emperors Diocletian and Constantine, and claim to use sources which range from acceptably solid Greek and Latin works, to apparently fake histories concocted in their own imaginations. This free-wheeling conduct on the part of the authors has therefore led to a series of works that range in quality from those of reasonably secure sources of historical fact, to fantasies which have their 'historical' sources in the fabrications of the authors. (Momigliano, 1954) The quality also ranges from plodding, but reasonable accounts of men such as Antoninus Pius or Marcus Aurelius (author of the 'Meditations'), to lengthy and scabrous descriptions of the activities of Heliogabalus, a reported pervert who exercised the absolute power of imperial rule for a number of years after 218 of this era. It even contains one so-called 'biography' that describes some 33 different usurpers or attempted usurpers, all in insufficient detail to permit an understanding of their careers.

The importance of the Historia Augusta is unde-

niable, since it is virtually the sole Latin text of any substantial length of that age, and can be matched only by sections of histories written in Greek. For some of the period, it is the only source, and for this reason, the question of authorship is indissolubly linked to the problem of disentangling any solid historical information that is contained in parts of the work, from the pure fantasy that infects too much of it.

The Historia Augusta as a whole comprises some 111,000 words, or, without personal names, some 101,000. The individual lives vary in length from the incomplete life of Valerian, with 1,030 words of text, of which 400 are dictionary head-words other than personal names, to the text of Alexander Severus, with 11,000 words of text, including 1,900 dictionary head-words.

Design Decision

The varying quality of the works, and the risks of making subjective judgements through statistical analysis based upon a limited number of critical words, brought us to act upon two decisions:

a) The complete text has been lemmatized and disambiguated so as to be available for stylometric analysis (by a system described at the ALLC/ACH conference in Paris in 1994). This produces a file of parallel lines of the original forms and of their lemmas, with all matching words and punctuation being vertically aligned.

b) This text is used to generate a file of word frequencies that provides an array (or, matrix) of approximately 5,600 disambiguated lemmas (dictionary head-words), together with their frequencies of occurrence in each of the thirty lives. At the discretion of the researcher, comparable files of the original forms can also be generated, and personal names can optionally be included in this matrix. The design also allows more than 20 other arrays to be generated; for example:

- frequencies of occurrence of specific lemmas, or classes of lemmas, per arbitrary number of input words;
- frequencies of only those lemmas that are represented in all 30 lives;
- identification of the biographies that use each specific lemma fewer than an arbitrary number of times (usually less than 30 occurrences overall in the H.A.);
- 30×30 (symmetric) matrices giving the numbers of lemmas unique to, and also common to (but not unique to), each pair of biographies;
- various frequencies of spacings between first appearances of lemmas (i.e.: the frequency of introduction of new lemmas);
- lengths of words and sentences; uses of the various punctuation marks and of the letters of the alphabet;
- frequencies of use of named function words, and of coordinating and subordinating conjunctions (for research especially in certain topics in psycho-linguistics);
- various word frequencies, suitable for the differing requirements of Zipf's Law and of Yule's Characteristic.

All arrays are produced also in 'comma-delimited' form, and can be transmitted directly to a database, spreadsheet, and to statistical packages such as SPSS or SAS.

Basic to the design of the research was the need to be able to take groups of segments of biographies for statistical analysis. Hence the entire system is under the control of a 'batch' file, so that once a researcher has chosen the segments of text to be analyzed (by use of a text editor), a single command results in the creation of the matrix or matrices of frequencies. In the case of the entire text of the 30 biographies, this requires 93 minutes on a 33 MHz DOS machine to produce the fundamental matrix of frequencies, or 5 hours for the complete set of files of matrices.

There was a further purpose behind this decision to create the matrices. The question of acceptable disambiguation is never closed, since many decisions are subjective, and are not accepted by all scholars. For example, if an adverb has been created from an adjective that has itself been created from a verb form, is the adverb to be lemmatized to its form as an adverb, or to the dictionary head-word of the original verb? Our design decision was made so that a scholar who disagrees with the lemmatization and disambiguation, or even with the text itself, and possibly with its punctuation, or with the lists of subordinating conjunctions or the function words, can modify the original files in any manner desired (using a text editor), then run the 'batch' file again, and develop a new set of matrices in just over five hours.

Statistical Analysis

Many different analytic measures, such as correspondence analysis, factor analysis, cluster analysis and principal component analysis, have been employed in this study. Principal component analysis, for example, has been used to examine the occurrences of various frequently used lemmas found in all 30 texts. The occurrences of words found in only some texts has also been examined, since this, too, is an element of style which may be used to examine authorship (Burrows, 1992).

Vocabulary richness has been examined in many ways. For example, the hapax/token ratio has been examined for segments of text of approximately

equal lengths, and Honoré's (1979) formula has also been used in the examination of texts in reference to hapax legomena and dislegomena.

The matrices provide an ease of use in these studies, in that the relative numbers of hapax legomena and dislegomena are available at a glance. Furthermore, utility programs (or simple counting, in shorter texts), can find lemmas and forms common to two or more texts (where forms, of course, may be ambiguous) as well as sets of words unique to any given text or to subsets of texts.

The use of the original or the lemmatized text, in close conjunction with the matrices, permits another valuable means of analysis of the richness of vocabulary. For the text itself permits analysis of the rate of introduction of new words, and of the degrees of separation between new words over the entire span of each of the texts being considered; but since each specific lemma or form has its own row in the matrix of frequencies, different rows can be combined for analysis of the use of synonyms of words discovered in the earlier analysis, and hence can provide a further measure of the vocabulary of an author. The lemmatized text can also, of course, be transmitted directly to other systems of analysis, such as TACT.

Further Considerations

It should be noted that modern texts, which may not require the intensive lemmatization and disambiguation that is necessary in highly inflected languages such as Latin, can be transmitted directly to the programming system that generates the various matrices of words and frequencies of occurrence.

Conclusion

The preponderance of evidence at present in our research points to a degree of multiple authorship (or of authors plus editor). Yet the question of authorship of the 'Historia Augusta' may never be solved to the satisfaction of all, and it probably cannot be solved by stylometric methods alone; but a judicious balance between stylometric and literary and historical techniques will improve our understanding of this interesting, controversial, and infuriating work. It will also assist in solving the fundamental problem of the Historia Augusta: the question whether its analysis of the lives of the emperors represents the frame-of-reference (and hence the interpretation) of a single author, or is the quite significantly less valuable result of perhaps a number of persons, writing over an unknown span of time.

Bibliography

Burrows, J.F. (1992). "Not Unless You Ask Nicely: The Interpretive Nexus Between Analysis and Information", Literary and Linguistic Computing, 7: 91-109.

Chastagnol, A. (1994). (Traducteur du latin de l'Histoire Auguste). Les Empereurs Romains des IIe et IIIe Siècles. Éditions Robert Laffont, S.A. Paris.

1991. Historia Augusta. Packard Humanities Institute CD-ROM 5.3.

Holmes, D. (1992). Authorship Attribution. A monograph produced for the Faculty of Computer Studies and Mathematics, University of the West of England, Bristol.

Holmes, D. (1994). "Authorship Attribution", Computers and the Humanities, 28: 87-106.

Honoré, A. (1979). "Some Simple Measures of Richness of Vocabulary", Association for Literary and Linguistic Computing Bulletin, 7: 172-177.

Momigliano, A. (1954). Journal of the Warburg and Courtauld Institutes, 17: 22-46. Re-printed in: "An unsolved problem of historical forgery: the 'Scriptores Historiae Augustae'", Studies in Historiography .(1969): 143-180.

Sichel, H. S. (1975). "On a Distribution Law for Word Frequencies", Journal of the American Statistical Association, 70:542-547.