Trends and Issues in Quantitative Stylistics

D. L. Clayman
Brooklyn College and The Graduate Center
The City University of New York

Recent publications by Greenberg, Frischer, and Brandwood suggest renewed interest in quantitative stylistics as a philological tool. Since we can expect more work in this vein, thanks to the TLG and PHI, which have given us so many reliable machine–readable texts, this seems like a good moment to consider briefly what has been achieved, what difficulties remain, and what the future could hold.

Both Greenberg and Frischer are working in traditional ways, traditional, that is, within the short history of modern stylometry. Greenberg's tabulations of word juncture in Latin prose and poetry belong to a European "school" with centers in Liège at Le Laboratoire d'analyse statistique des langues anciennes (LASLA) and in Tübingen at Wilhelm Ott's Zentrum für Datenverarbeitung.[1] These centers specialize in data generation. Various features of whole texts, such as phoneme distribution, types of subordination, or an author's lexicon, are counted, tabulated, and the results published with minimal statistical analysis. No special justification is offered for the choice of items counted, no hypothesis is proved or disproved in the process, and no claims are made for the results except the hope that they will benefit future work. Delatte and his collaborators, for example, have studied subordination in 15 Latin authors with a corpus of texts totalling about 800,000 words; their only conclusion is that the ancient authors knew 151 possible kinds of subordination among them, but none uses more than half of the total. The scholarly objective is simply to establish a baseline of information about literary language with which new data can be compared so what is stylistically unique will not be confused with what is ordinary. We cannot define linguistic features that are specifically Vergilian, for example, unless we can discount what is usual in Latin and in Latin hexameters. This is true, of course, in traditional literary criticism, where wide reading provides the necessary baseline. When texts are measured rather than read, this purpose can only be served by other measurements.

Although the collection of quantitative information about classical texts has been growing steadily since the mid–60s, more quickly in Latin than in

[1] LASLA's guiding spirit is Louis Delatte. Its principal organs of dissemination are its own Travaux and journal, generally called *Revue*, but *RELO* in *l'Année philologique*. The metrical analyses that are Tübingen's specialty appear in their series Materialen zu Metrik und Stilistik, e.g. Ott cited below.

Greek, there are still more gaps than useful contributions because initially the business of preparing texts and writing programs was slow and difficult. That is now largely behind us, and we can look forward to rapid gains in data generation. The value of a comprehensive baseline of information, especially at the lexical level, has been convincingly demonstrated by the success of L'Institut national de la langue française (INaLF) at Nancy where the machine–readable texts of Le Trésor de la langue française have been systematically analyzed with results published mainly in the series TQL (Travaux de Linguistique Quantitative). Among these, the most impressive is Étienne Brunet's *Le Vocabulaire français de 1789 à nos jours*, a massive analysis of French literary language spanning more than 350 authors and 1,000 texts with 70,000,000 words. Nothing like this in scale or value has ever been produced in classics, but it is the kind of broadly based analysis that we can look forward to in the future and can count on to generate new knowledge about Latin and Greek.

The analyses that have flowed from Nancy are greatly enhanced by the systematic application of statistical measures, pioneered by Charles Muller (and reviewed critically by D. Ratkowsky, *CHum* 22 (1988) 77–85), that facilitate comparisons of individual authors, genres and periods. LASLA, in contrast, has shown greater caution, generally preferring tables to formulas, and data collection to interpretation. Yet, it is statistical analysis that gives data value and meaning. Statistics have the power to summarize large numbers of discrete facts and to discover relationships among them. They make possible precise comparisons between texts divided or grouped in any fashion; they can prove that relations between them exist that may not be apparent to even a careful reader, and they do all of these following uniform standards universally accepted in the real world. Although classicists have not traditionally been trained in their use, nothing prevents us from mastering statistics except our own anxieties.

But what statistics are appropriate for literary data? No question is more vexed. The nub of the difficulty is that the most common statistical techniques, developed for the sciences and social sciences, make assumptions about data that are not true for languages or texts. Textual phenomena are neither random nor normally distributed (Fortier), though most common statistical tests assume that they are. Even non–parametric, distribution–free statistical measures like chi–square are severely compromised by lack of independence in the data (Usher & Najock, Stevenson). What we need is a mathematical model designed to suit linguistic structures, taking account of their unique nature, which is highly structured, diachronic and interdependent. No such thing yet

exists, and until it does, we will have to learn how to choose wisely from among the statistical techniques available.[2] Discourse, cluster and time-series analysis have all been used effectively with textual data and there are other possibilities as well. Good commercial software has eliminated the difficulty of doing the calculations, but it cannot solve the thornier problems of matching the technique to the data or drawing reasonable conclusions from the results.

The intelligent choice and handling of statistical measures depends on two sets of assumptions: assumptions about the data made by each statistical test and the assumptions about language, literature, and the act of writing that underlie the philological or historical point at issue. The importance of the latter is illustrated clearly by chronological and authorship studies. These were the earliest stylometric applications in literary studies—they go back at least to Campbell—and remain the most problematic. They proceed in a way exactly opposite to LASLA, beginning with a thesis and admitting only data and statistical analyses that appear to prove it. It is not the thesis *per se* that makes these studies problematic, but the assumptions on which the quantitative arguments are based. Quantitative approaches to both authorship and chronological problems assume that the features counted are independent of the content of the work, the genre, intertextuality, and the author's will, which simply cannot be true. In addition, authorship studies assume that an author's style is as personal and definable as a fingerprint, with essentials unchanging through time, while chronological studies assume the opposite, that an author's style is evolutionary, changing in a consistent way over a lifetime of writing. No one has proven conclusively that either is true, as a general rule, and placed side by side, they contradict each other.

Unlike many of his predecessors in this area, Frischer is sensitive to the assumptions that underlie his methodology. His use of context-free function words to date the *Ars Poetica* (a technique developed by Mosteller and Wallace) and his proofs that their frequency changes through time according to definable trends in Horace's other, datable poems, while controlling for meter, demonstrate an understanding of the issues that argues for confidence in his results.[3] It was not possible, of course, for Brandwood and the others who have

---

[2] The breakthrough will most likely come from the sciences. Recently, cladistic analysis, developed by evolutionary biologists, has shown real promise as a technique for determining manuscript stemmata: P. Robinson and R. J. O'Hara, *BMCR* 3 (1992).

[3] For an opposing view, see Keyser's review of Frischer in *BMCR* 3 (1992) 118–122; cf. his review of Brandwood, *BMCR* 3 (1992) 58–74. Keyser's complaints are similar in tone to those repeatedly made by W. A. Smith about the authorship studies of A. Q. Morton, who is in turn defended by Merriam (and attacked in turn by Smith, *CHum* 21.1 [1987] 59–60). The dispute eventually inspired Thomson's "How to Read Articles that Depend on Statistics."

worked on chronological problems in Plato to validate their criteria in the same way because none of the dialogues has a secure date. Like all areas of classical studies, quantitative analysis suffers badly from gaps in the evidence.

However attentive one is to methodology and however secure the facts from which the argument proceeds, chronology and authorship studies are often viewed with skepticism. In an effort to produce more convincing conclusions, more and different criteria have been counted and subjected to statistical tests of increasing complexity. Thus, Craik and Kaverly look at the distribution of all 280,413 letters in the Oxford text of Sophocles and use principal component analysis to date the *Trachiniae* early, A. B. Nicolova addresses the relative chronology of three essays of Seneca by counting figures of speech, while Usher and Najock tabulate frequent words, 13 word classes (articles, adjectives, adverbs, participles, etc.), and total vocabulary distribution, on which they unleash a variety of statistical techniques to identify spurious speeches in the *Corpus Lysiacum*. In a rare display of scholarly candor, they conclude their paper by disagreeing with each other over the interpretation of the results, illustrating all too clearly the difficulty of using inferential judgments in a philological argument (103–04). The most recent scholar to deploy massed data in an authorship study is M. Lana, who uses correspondence analysis to marshal the total vocabulary distribution of Xenophon, Thucydides, and Aristotle's *Ath. Pol.* and *Pol.* with a view to isolating Xenophon's *Ath. Pol.* and *Lak. Pol.* from his *HG*, to prove their false attribution. At this point, authorship and chronology studies begin to rival LASLA as purveyors of baseline information, and the two approaches converge.

Though data collection is important *per se*, one has to admit that data vivified by a hypothesis is much more interesting. Hypotheses about authorship and chronology, however, are by no means the only kind amenable to a quantitative approach. Most classicists, like literary critics in any language, will more likely be interested in quantitative studies like Potter's "Character Definition through Syntax," Ide's study of images in Blake's *The Four Zoas*, and Burrows's examination of how Jane Austin's characters evolve linguistically through narrative time. Classicists have also entered these waters with Thury's investigation of the images of youth and old age in Euripides *Supplices*, Hubka's analysis of topological structures in new comedy, Delatte's study of style and character in *Heroides* xvi and xvii, and Sale's quantitative criteria for determining the relative age of Homeric formulae.

When quantitative analysis tackles subjects like character, dialogue, image, and structure it becomes far more relevant to what most scholars care

about and begins to answer W. van Peer's devastating criticism: "While devoting all energy to the lower levels of language organization, at the expense of the textual dimension, quantitative studies sacrifice not only the cultural significance of literary texts, but also the generalizability of its own findings" (305). His own list of aspects of textual organization that could be fruitfully quantified includes text deixis, turn-taking mechanisms in dialogue, speech and thought presentation, parallelism, alternations in point of view, openings and closings, plot and the use of 'editorial' techniques by a narrator (306).

What is needed is a link between quantitative stylistics and mainstream literary theory. In an essay that should be read by anyone contemplating work in this field, John Smith has shown how this bridge can be built by establishing a common ground between what he calls "computer criticism" and Formalist/Structuralist perspectives on text. The computer's ability to map literary phenomena both synchronically and diachronically creates the potential for adding rules of evidence, strict concepts of proof and a discovery procedure for interpretive generalizations to a "soft" theory lacking all three. In short, Smith shows how computer criticism could give literary criticism the rigor and explanatory power of science.

Though its past performance did not live up to the impossible expectations that were once held out for it (note the laments in *CHum* 25.6 [1991]), quantitative stylistics offers more possibilities for fruitful work than most classicists know. Like colleagues setting out to do more common sorts of literary criticism, those who wish to pursue it should read widely in the literature—Potter's anthology is a good place to start—and should not limit their reading to previous work in classics.

Works Cited

(TLG is the Thesaurus Linguae Graecae at the University of California, Irvine, directed by Theodore Brunner. PHI is the Packard Humanities Institute of Los Altos, CA founded by David W. Packard.)

Brandwood, Leonard. *The Chronology of Plato's Dialogues* (Cambridge 1990).

Brunet, Et. *Le Vocabulaire français de 1789 à nos jours*, 3 vols., TQL 17 (Geneva and Paris 1981).

Burrows, J. F. *Computation into Criticism: A Study of Jane Austin's Novels and an Experiment in Method* (Oxford 1987).

Campbell, L. *The Sophistes and Politicus of Plato* (Oxford 1867).

Craik, E.M. and D. H. A. Kaferly. "The Computer and Sophocles' *Trachiniae*," *LLC* 2,2 (1987) 86–97.

Delatte, Louis. "Recherches statistique sur les *Heroides* xvi et xvii d'Ovide," *Revue* 1979,2 1–61.

_____, Et. Evrard, S. Govaerts & J. Denooz. *Dictionnaire fréquentiel et index inverse de la langue latine* (Liège 1981)

_____,S. Govaerts, J. Denooz. "La subordination en Latin," *Revue* 1981,1–4 1–195.

Fortier, P. A. "Theory, Methods and Applications: Some Examples in French Literature," *LLC* 6.3 (1991) 192–193.

Frischer, Bernard. *Shifting Paradigms: New Approaches to Horace's Ars Poetica*, American Classical Studies 27 (Atlanta 1991).

Greenberg, Nathan A. "Word Juncture in Latin Prose and Poetry," *TAPA* 121 (1991) 297–333.

Hubka, K. P. "Scenic Dominance, Motif and Conflict in the Topological Structures of the New Comedy," *ALLC Bulletin* 13 (1985) 31–36 & 61–66.

Ide, N. M. "Patterns of Imagery in William Blake's *The Four Zoas*," *Blake: An Illustrated Quarterly* 20.4 (1987).

Lana, M. "Xenophon's *Athenaion Politaeia*: A study by Correspondence Analysis," *LLC* 7.1 (1992) 17–26.

Merriam, Thomas. "An Investigation of Morton's Method: A Reply," *CHum* 21.1 (1987) 57–58.

Morton, A. Q. *Literary Detection* (Bath 1978).

Mosteller F. and D. L. Wallace. *Inference and Disputed Authorship: The Federalist Papers* (Reading MA 1964).

Muller, C. *Principes et méthodes de statistique lexicale* (Paris 1977).
Initiation à la statistique linguistique (Paris 1968).

Nicolova, A. B. "De la brieveté de la vie de Sénèque. Essai de chronologie par stylométrie," *Revue* 22 (1986) 95–103.

Ott, W. *Metrische Analysen zu Lucrez, De Rerum Natura, Buch I*, Materialen zu Metrik und Stylistik 6 (Tübingen 1974).

Potter, R. G. "Character Definition through Syntax: Significant Within–Play Variability in 21 Modern English–Language Plays," *Style* 15.4 (1981) 415–434.
ed. *Literary Computing and Literary Criticism* (Philadelphia 1989).

Sale, W. M. "The Formularity of the Place Phrases of the *Iliad*," *TAPA* 117 (1987) 21–50.

Smith, John. "Computer Criticism," *Style* 12 (1978) 326–56, rpt. in Potter (1989) 13–44.

Smith, M. W. A. "An Investigation of the Basis of Morton's Method for the Determination of Authorship," *Style* 19.3 (1985) 341–68.

_____,"An Investigation of Morton's Method to Distinguish Elizabethan Playwrights," *CHum* 19 (1985) 3–21.

_____, "The Revenger's Tragedy: The Derivation and Interpretation of Statistical Results for Resolving Disputed Authorship," *CHum* 21,1 1987 21–55.

_____, "Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship," *LLC* 2,3 (1987) 145–152.

Stevenson, B. "Adapting Hypothesis Testing to a Literary Problem," in *Literary Computing and Literary Criticism*, ed. R. Potter, (Philadelphia 1989) 65–68.

Thomson, Norman. "How to Read Articles which Depend on Statistics," *LLC* 4,1 (1989) 6–11.

Thury, E. M. "A Study of Words Relating to Youth and Old Age in the Plays of Euripides and Its Special Implications for Euripides' *Suppliant Women*," *CHum* 22,4 (1988) 293–306.

Usher S. and D. Najock. "A Statistical Study of Authorship in the Corpus Lysiacum," *CHum* 16 (1982) 85–105

van Peer, W. "Quantitative Studies of Literature. A Critique and an Outlook," *Proceedings of the Eighth International Conference on Computers and the Humanities*, *CHum* 23,4–5 (1989) 301–07.