Title	Beyond Lagado: An outlook for computational stylistics
Sub Title	ラガドを超えて : コンピュータによる文体分析の展望
Author	Armour, Andrew
Publisher	慶應義塾大学藝文学会
Publication year	1990
Jtitle	藝文研究 (The geibun-kenkyu : journal of arts and letters). Vol.58, (1990. 11) ,p.221(168)- 226(163)
JaLC DOI	
Abstract	
Notes	慶應義塾大学部文学科開設百年記念論文集
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00072643-00580001-0226

Beyond Lagado: An Outlook for Computational Stylistics

Andrew Armour

After descending from the flying island of Laputa, Gulliver came one day to the grand academy of Lagado where he met with a "projector", a professor in speculative learning who had built a large "frame" of wood, wire and paper, with which he hoped to write great books, simply by turning handles to create random arrangements of words:

> He assured me that this invention had employed all his thoughts from his youth; that he had emptied the whole vocabulary into his frame, and made the strictest computation of the general proportion there is in books between the numbers of particles, nouns, and verbs, and other parts of speech.[1]

In having Gulliver heap praise on "this illustrious person" and his "wondrous machine", Swift leaves no doubt as to his own mistrust of those who subject works of literature to the "strictest computation". Although, after two and a half centuries, the skepticism has changed little, the technology is a great deal more sophisticated and we can now look forward to rapid progress in the field of computational stylistics.

At the heart of Swift's mistrust lies the belief that "style" is too nebulous a concept to be accessible to rigorous analysis. Indeed, it is considered by some to be as elusive — and, for practical purposes, as useless — as the physicist's ether.[2] Yet this has not prevented men such as Swift and Buffon from offering definitions that serve to foster

rather than foreclose debate.

It is Buffon's famous contribution — "Style is the man himself" — that expresses the basic premise behind those authorship studies that take a statistical approach to stylistics: namely, a writer has a distinctive and objectively describable style. There are, of course, complications presented by a writer's determined efforts to imitate another's style, as in a pastiche, or to impose unnatural limitations on the process of composition, as in a lipogram.[3] Nevertheless, the stylistician works on the assumption that it should be possible to distinguish between the works of two writers, provided that idiosyncratic elements of style can be identified. Herein lies the problem.

Faced with the rich choice of lexical, grammatical, syntactical and rhetorical possibilities offered by a language such as English, the writer inevitably makes decisions about "how" as well as "what" to communicate. Some of these decisions will be conscious, some not; the latter are obviously prized by the stylistician. And working on the hypothesis that these decisions are not random — that patterns will be observed in the works of a particular author or specific genre — the hunt is on for characteristic idiolects or word habits that will help to discriminate one writer from another.

Clearly certain stylistic changes would be expected to result if the writer changes topic, or if the intended readership is different, or the writer has simply matured. The researcher thus looks for stylistic criteria that are (a) non-contextual, and (b) consistent. T. C. Mendenhall, a pioneer in this field, chose word-length frequency distributions — which he called "characteristic curves of composition" — to investigate the works of Dickens, Thackeray, Shakespeare and Bacon.[4] Although simplistic by today's standards, this method was based on the plausible notion that a writer's active vocabulary is to a certain degree unique. Unfortunately, Mendenhall had to admit that the curves of composition he plotted readily reflected any conscious effort on the part of the writer to mimic a certain style.[5]

Sentence length was the discriminator chosen by Udny Yule for his 1938 work on *De Imitatione Christi*,[6] in which he concluded that

Thomas à Kempis, not Gerson, was the author. He later tackled the same problem with a new yardstick: the frequency distribution of nouns.[7]

It was at about the same time a statistician named Frederick decided to apply numerical methods to the *Federalist* papers, a collection of anonymous essays written in 1787 and 1788. It is known that three men were responsible—Alexander Hamilton, James Madison and John Jay. In fact Hamilton provided a list indicating the authorship of each paper, but this was later challenged. Of the 77 papers, 12 were claimed by both Hamilton and Madison.

Making use of the undisputed papers, Mosteller calculated sentence length, as well as percentages of nouns, adjectives, one- and two-letter words, and the definite article. Although his results suggested that Madison was the author in most cases, he admitted that the method was not sufficiently sensitive. Hamilton's and Madison's prose styles turned out to be unusually similar as regards average sentence length, and Mendenhall's curves of composition proved to be totally ineffective as a means of discriminating between them. Obviously more reliable stylistic features had be found.

Mosteller returned to the problem about twenty years later, after Douglass Adair informed him that he could distinguish the styles of Hamilton and Madison on the basis of a "proportionate pair" of marker words — the former preferring *while* and the latter *whilst*.[8] In collaboration with David Wallace, Mosteller then undertook four studies and published the results in 1964: their conclusion was that all twelve papers were indeed written by Madison.[9]

On the face of it, an investigation of low-frequency words which have been shown by screening to be used unevenly between the two writers in question would seem to be a promising line of enquiry. However, in the case of the *Federalist* papers, the strongest evidence came from a group of 8 high-frequency function words: *also, an, by, of, on, there, this, to.* Such words are least likely to be context-bound. In fact it is for this very reason that most are included in stop lists of "noise words" to be excluded during automatic indexing. What is

discarded by the indexer is potentially very valuable to the stylistician. A similar method was adopted recently by Wilfred Smith in his study of *Pericles*, enabling him to conclude that Wilkins, not Shakespeare, probably wrote Acts I and II.[10]

The numerical analysis of one or two features of literary style can yield interesting data, but this is only scratching the surface. It is thus not surprising that the publication of such results — convincing though the researcher may find them — is often greeted with the sort of polite comments made by Gulliver. But then pioneers such as Mendenhall had little more to aid them in their research than the wooden frame used by the projector of Lagado. Fortunately, however, the situation is about to change, thanks to the advent of the computer.

Today's stylisticians have the freedom to study literary works with methods that only a few decades ago would have seemed impossibly tedious. Of particular importance is the fact that the number of stylistic criteria can be greatly increased. For instance, Thomas Merriam investigated 41 "word habits" in *The Booke of Sir Thomas More*, concluding that it was written solely by Shakespeare.[11] Similarly, Yehuda Radday employed 56 criteria of "language behavior" to conclude that *Genesis* is probably the work of a single author.[12] And more recently, Anthony Kenny examined 96 features in a search for stylistic differences in the Pauline Epistles.[13] Theoretically there is no limit to the number of stylistic features than can be studied, but among those used in past studies are:

- type/token ration and their inverse ("pace")
- Yule's K index (an inverse measure of richness of vocabulary)
- frequency profiles based on the number of syllables per word
- patterns of coordination and subordination
- word links and the percentage of conjunctions
- the types and depths of "nesting" in sentences.
- sentence-initial structures
- the position of certain words — especially *hapax legomena* (once appearing words) — within sentences

- the use of metaphor and simile
- the use of comparative and superlative forms
- conditional clauses and phrases
- methods of enumeration
- punctuation

Naturally, many of these can be further combined, as in noun-adjective and verb-adjective ratios. The computer also makes possible multivariate analyses — such as cluster analysis — that take into account several features simultaneously.

Now that pioneering studies, such as that of Mosteller and Wallace, have helped to identify the more reliable measures of an author's style, it is time for the piecemeal approach to computational stylistics to be abandoned in favor of a multilevel analysis of texts. Once a consensus is reached on text encoding standards, it should be possible to develop software that can follow established statistical procedures for homogeneity, etc., unattended. For example, the chi-square statistic and coefficients of correlation would be used to determine whether an observed disparity is statistically significant; if so, it can be flagged and brought to the attention of the researcher. Furthermore, methods should be found for the simultaneous representation of the most important stylistic features of a text — something akin to a weather map, perhaps — to facilitate quick comparisons.

Part of this proposed package would be period- and genre-specific dictionaries that would enable the software to, say, distinguish content from function words, or perform simple content analysis operations. Semi-automatic lemmatization and disambiguation should be possible using techniques derived from AI and machine translation research. We may even hope for fuzzy mathematics to make a contribution.

Seen in perspective, this is clearly an emerging field, and one in which exaggerated claims are sometimes made by those with more enthusiasm than exactness. And it should not be forgotten that, however sophisticated the methodology or the technology may become, firm "proof" can never be furnished — only probabilities that serve to

strengthen or weaken the subjective hunches of the researcher. But by adopting a holistic approach to the numerical analysis of literary texts, it will soon be possible to provide the Gulliver skeptics with far more convincing evidence for the worth of computational stylistics as an alternative to traditional forms of literary research.

Notes

1) Jonathan Swift, *Gulliver's Travels* (1726; rpt. London: J. M. Dent, 1912), p.169.
2) Bennison Gray, *Style: The Problem and its Solution,* as quoted by G. W. Turner in *Stylistics* (Harmondsworth: Penguin, 1984), p.238.
3) A lipogram is a work written in such a way as to avoid the use of specific letters of the alphabet. Two notable examples are Ernest Vincent Wright's *Gadsby* (1939) and George Perec's *La Disparation* (1969), both of which omit the letter "e".
4) T. C. Mendenhall, "The Characteristic Curves of Composition," *Science,* 9.214, supplement (March 1887), pp.237-49; also, "A Mechanical Solution of a Literary Problem," *Popular Science Monthly,* 60.2 (December 1901), pp.97-105.
5) "A Mechanical Solution of a Literary Problem," p.105.
6) Udny Yule, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship," *Biometrika,* 30 (January 1938), pp.363-90.
7) Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge: Cambridge University Press, 1944).
8) Adair provided information in 1959, according to Ivor S. Francis, "An Exposition of a Statistical Approach to the *Federalist* Dispute," *The Computer and Literary Style* (Kent, Ohio: Kent State University Press, 1966) p.40.
9) Frederick Mosteller and David L. Wallace, *Inference and Disputed Authorship: The Federalist* (Reading, Mass.: Addison-Wesley, 1964).
10) M. W. A. Smith, "The Authorship of *Pericles*: New Evidence for Wilkins," *Literary and Linguistic Computing,* 2.4 (1987), pp.221-230.
11) Nigel Hawkes, "Computer Finds 'New' Play by Shakespeare," *The Observer,* 6 July 1980, p.1.
12) "By One Hand?" *Time,* 7 December 1981, p.42.
13) Anthony Kenny, *A Stylometric Study of the New Testament* (Oxford: Clarendon Press, 1986).