

Statistics in Language Studies: Neglect, Use, and Abuse

Jane Buck
Department of Psychology
Delaware State University

A paper presented at the 38th annual conference of the International Linguistics Association, New York City, April 1993.

Mosteller and Wallace (1984) introduced their study of statistical methods applied to a question of disputed authorship by saying, "When two statisticians, both flanks unguarded, blunder into an historical and literary controversy, merciless slaughter is imminent." (p. 1) Most of my formal training, in contrast, is not in statistics, but in the social and behavioral sciences. I came slowly and late to an appreciation of the utility of statistics in a variety of disciplines, including my own. Through colleagues in linguistics, classics, and literary studies, I have recently become aware of some of the problems language scholars are attacking that would be more easily solved if approached statistically.

My first impression of much current work in linguistics was that very little use has been made of statistical methods. However, as I delved more deeply into the topic, I discovered that there is a venerable, if uneven, history of the application of statistics in language studies. Time allows only the most superficial and cursory glance at this history.

I shall proceed chronologically, for the most part. This approach allows for a fairly tidy topical division, as well. Many early studies using quantitative methods were concerned with issues of style, either for the purpose of establishing authorship or of dating manuscripts by a known author. More recently, linguists have also used quantitative methods in historical linguistics and in the attempt to discover language universals. Psycholinguists, who straddle the line between psychology and linguistics, have been more eager than their colleagues in other branches of the field to embrace statistics as part of their tool kit.

If we expand the definitions of linguistics and statistics to include quantitative language studies undertaken before linguistics and statistics acquired names, the history of statistics in linguistics can be traced to the work of the medieval Masoretes, who were committed to preserving the purity of Hebrew texts. Much of their work, which consisted of counting virtually everything that could be counted, is of little interest to language scholars, and was probably undertaken to determine the compensation of copyists, who were paid by the letter (Yule, 1968).

The term "statistics" originally meant a collection of data referring to states, in other words, political economy. Almost two hundred years have passed since William Playfair, generally acknowledged, albeit erroneously (Biderman, 1990), as the father of graphical statistics, developed his methods. Since then, "statistics" has come to mean, at least to statisticians, the analysis of data for the purpose of description and, more importantly, of making inferences from limited data. What the non-statistician calls "vital statistics" should, in the contemporary statistician's view, be called "vital data"; and what the non-statistician conceives to be statistics might better be characterized as data collection or enumeration.

One of the earliest suggestions for the application of statistics in the modern sense to a

linguistic problem was made in 1851 by Augustus de Morgan, who proposed in a letter to a friend that word length might be used to establish the authorship of disputed texts (Lord, 1958). However, there is no evidence that de Morgan, a mathematician, undertook any of the studies he suggested (Williams, 1956). Thirty-six years later, Thomas Corwin Mendenhall, a physicist, published a paper in which he cited de Morgan's notion. Mendenhall expanded and improved upon de Morgan's idea by comparing not only average word length, but the entire frequency distributions of word length (Williams, 1956).

Mendenhall believed that the comparison of word length distributions could be used to establish authorship, much as a metallurgist uses spectrograms to identify metals. Not surprisingly, he found very similar distribution curves for works by the same author and markedly different curves for samples of Latin, Italian, and German. Unfortunately, the method did not prove to be particularly useful over a wide range of problems. In a comparison of the works of Shakespeare and Marlowe, the curves of the two authors were virtually indistinguishable (Williams, 1956). It is not surprising, of course, that word length failed to discriminate between authors in a reliable fashion, in that the choice of word length is limited by the vocabulary available in a given language. Over 50 years later, G. Udny Yule, a noted statistician, published an article in *Biometrika*, in which he proposed a comparison of sentence length as an alternative (Williams, 1956). Sentence length, of course, is, to a much greater degree, under the control of the individual than is word length, and thus appeared to have more utility as a discriminator.

In 1944, Yule published a germinal study of the disputed authorship of *De Imitatione Christi*, traditionally attributed to Thomas à Kempis, although there were several other candidates (Yule, 1968). Yule was motivated, not by an imperialistic desire to impose his discipline on language scholarship, but by an interest in the work itself. (Ironically, the author of the *Imitatio* counseled the reader not to ask who wrote the work, but to pay attention to what was written.) There had been a number of prior attempts to determine whether the *Imitatio* was the work of Thomas or of one Jean Charlier de Gerson, all of which Yule found inadequate to the task. They focused on extraneous details and thus failed to give an adequate description of the work as a whole. In Yule's words:

To tell me that there is a small mole on Miranda's cheek may help me to identify the lady, and may in conceivable circumstances be quite useful information to the police, but it hardly amounts to a description of her alluring features. (p.2)

Yule brought to the enterprise, by his own apologetic admission, an ignorance of linguistics that was balanced, not only by a passionate love of words and a scholarly curiosity, but by a mastery of over half a century's progress in statistical methodology, including advances in sampling theory, correlational techniques, tests of goodness of fit, and the ability to quantify variability. Early in his investigation, he struggled with a number of problems. What linguistic unit would best serve as a discriminator? Having settled on the word as the appropriate unit, what part(s) of speech should be examined? Should the investigation concentrate on rare or commonly used words? How large a sample of each author's works would be required? Are characteristics of language stable over various sample sizes (Yule, 1968)? Consequently, Yule devoted seven of eleven chapters to methodological and statistical issues, confining his analysis of the *Imitatio* to Chapters 9 and 10. Many of the

methodological problems he attacked and the solutions he proffered, although perhaps naïve from a linguist's perspective, continued to require the attention of both statisticians and linguists for many years (Bailey, 1969). Yule reasoned that the choice of vocabulary, reflected in the frequency with which the two authors used given words, would serve to discriminate their work. Based on his analysis of the relative frequency with which the two candidates employed nouns, Yule concluded that the evidence weighed heavily in favor of Thomas à Kempis's authorship (Yule, 1968).

A number of refinements have been made to Yule's methods by other statisticians interested in stylistics (Bailey, 1969). One of the most frequently cited studies is another pioneering work involving statistical approaches to a case of uncertain authorship (Mosteller, 1984). In 1964, Frederick Mosteller and David L. Wallace published *Inference and Disputed Authorship: "The Federalist,"* which was reprinted 20 years later as *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. The *Federalist*, first published in the late eighteenth century, consists of 85 short essays, written anonymously by Alexander Hamilton, James Madison, and John Jay. The identity of the author or authors had been established, to the satisfaction of most historians, for the bulk of the papers. The problem was that of attempting to determine whether Madison or Hamilton wrote 12 essays whose authorship was uncertain (Mosteller, 1984).

Unlike Yule, Mosteller and Wallace were motivated, not so much by an intrinsic interest in the material as in contributing to statistical knowledge in the area of classification. They viewed the authorship problem as a case study that would allow them to compare two statistical methods—the so-called classical method developed by R. A. Fisher in the 1930's and the Bayesian method, named for the eighteenth century mathematician, Thomas Bayes, on whose theorem the method is based. They concluded, along with historians who had previously studied the problem, that all but one of the disputed papers were written by Madison, that one exception remaining in doubt (Mosteller, 1984).

Other statistical studies of disputed authorship include the novel *And Quiet Flows the Don*, investigated by Medvedev; editorials in the Indian newspaper *Kesari*; the German romantic work, *Die Nachtwachen*; nineteenth century papers on economics; numerous scriptural questions; and forensic disputes involving forged wills and faked testimony (Mosteller, 1984). Other fertile areas of investigation include chronology problems and diachronic stylistics. One such study cited by Mosteller and Wallace is an analysis by Kenny of two texts thought to have been written by Aristotle at different periods in his life: the *Nicomachean Ethics* and the *Eudemian Ethics* (Mosteller, 1984). Others, reported by Bailey (1969) in his historical survey, include the dating of Plato's works and of Spenser's *Cantos of Mutabilitie*, and Josephine Miles's characterization of English and American literature in terms of changing modes of sentence structure over several centuries. Bailey also cited a number of attempts to characterize individual styles in both prose and poetry. As a psychologist, I am pleased to report that B. F. Skinner published a study of sound-patterning in poetry, in which he compared poetic devices used by Swinburne and Shakespeare and developed a measure he called the coefficient of alliteration (Skinner, 1941).

Most of the studies I have cited so far have been the work of statisticians or other non-linguists. Language scholars have tended either to avoid using statistical methods in their work or to misuse them. Why? One historian of statistics (Royston, 1956) quoted William Playfair, as follows:

...for no study is less alluring or more dry and tedious than statistics, unless the mind and imagination are set to work or that the person studying is particularly interested in the subject; which is seldom the case with young men in any rank in life. (p. 245)

As one who teaches statistics to reluctant and often terrified students, I can attest to the accuracy of Playfair's observation, adding that it applies not only to young men, but to older men, and women of all ages, as well. Many linguists, along with other humanists and social scientists, are phobic with respect to quantitative methods. I have friends in language departments who actually take pride in their "innumeracy," perhaps as a defense against learning what they believe to be an insuperably difficult subject. Some language scholars genuinely believe that statistical methods are inappropriate to the study of language (Embleton, 1986). However, a growing number of linguists and other language scholars have recently begun to feel the need to quantify their data, but without a grounding in statistics have, on occasion, created more confusion than enlightenment. The confusion ranges from the misuse of statistical vocabulary to the inappropriate application of methodology.

In one study of topic continuity, for example, Derbyshire (1986) used the term "statistical significance" in a way that suggests that what he meant was that he had a sufficient number of tokens to analyze in a meaningful way. "Statistical significance" has a precise and unique meaning in statistics; it means that results are unlikely to have been caused by random factors. A probability value, such as .05 or .01 always accompanies a statement that the results are statistically significant. This number is obtained by subjecting the data to an appropriate test of significance and quantifies the probability that the results were caused by chance or sampling error. It says nothing about the importance or practical significance of the results.

In the same study referred to above, the author quite appropriately defined referential distance as the number of clauses between a referent and its last previous occurrence (Derbyshire, 1986). Referential distances of 20 or more clauses were arbitrarily designated as 20, as were newly introduced subjects. As one might predict, in the vast majority of instances, regardless of the way in which subjects were marked, topics had referential distances of one clause, with longer referential distances decreasing in frequency at a rapid rate. Thus, the distributions were markedly skewed. In such a situation, the arithmetic mean is seriously distorted by the extreme scores and should not be calculated. Further, it is mathematically impossible to calculate the mean when one has an open-ended category, such as "20 or more", because the actual values of all the data are unknown.

Without subjecting his data to a test of significance, the author stated that one way of marking subject was clearly the lowest in topic continuity, with a mean referential distance of 8.14, the second lowest having a mean referential distance of 5.11, and the highest, a mean of 1.31 (Derbyshire, 1986). The first category had only 22 tokens compared with 99 in the second, and 298 in the third. This enormous disparity in sample size, as well as the inappropriate use of the mean, is problematic from a statistical point of view.

There are ways of handling skewed distributions in a statistically sound manner; one could use the median rather than the mean or transform the data so that the distribution becomes symmetrical. I calculated the medians as one for the first two categories and 1.5 for the third. However, if, as I suspect, many, if not most, of the seven subjects included in the

20-or-more category were newly introduced topics, they should have been included in a category labeled zero or excluded from the analysis. If only one or two tokens were mistakenly categorized, the median would be one instead of 1.5, resulting in identical medians for all three categories. Derbyshire did, however, recognize that more than 50% of the cases for all methods of marking had referential distances of one, leading him to state that "...no single measurement is sufficient for establishing degree of topicality..." (Derbyshire, 1986, p. 254).

My purpose in pointing out some of the statistical shortcomings of the study is not to attempt to diminish the value of the work, but to encourage a more informed approach to the quantitative study of the topic. There is no need to re-invent the method, nor even to go outside the discipline for guidance. Isidore Dyen (1975) has published widely in the area of genetic lexicostatistics. His work demonstrates a sophisticated command of statistics, and of the way in which it can supplement traditional methods. He pointed out as early as 1960 that statistics had been used "...(1) to establish the relationship between languages, (2) to classify (subgroup) related languages, and (3) to establish the times at which related languages began to diverge" (p.75).

In her 1986 book, *Statistics in Historical Linguistics*, Embleton listed three areas in which statistics can serve the historical linguist: (1) as an adjunct to more traditional methods in establishing provisional family trees, (2) as objective evidence reinforcing the results obtained by other methods, and (3) "...to provide family trees...for language families for which there is no other means [emphasis hers] of determining a family tree..." (Embleton, 1986, p. 169).

I co-authored a paper with a linguist who has been working on the reconstruction of the Proto-Algic and Proto-Algonquian languages (Buck & Proulx, 1991). We explored the utility of statistical methods in establishing the existence or non-existence of a term in a proto-language when cognate terms do not exist in two or more daughter languages, as well as some methodological problems involving the use of extremely limited evidence.

Bernard Frischer (1991), in a study of Horace's *Ars Poetica*, used a number of modern statistical techniques to establish the date of the poem, a task that had proven extremely difficult for classical scholars using traditional methods. Parenthetically, he found that word length was not helpful, as did Mendenhall a century ago.

Sophisticated statistical methods have been available since the early part of the century and have been used with varying degrees of enthusiasm in language studies. I conclude with Embleton's exhortation: "Statistical methods should be used wherever possible in conjunction with or followed up by the more traditional methods" (Embleton, 1986, p. 169).

References

- Bailey, R. W. (1969). Statistics and style: A historical survey. In L. Dole el & R. W. Bailey (Eds.), *Statistics and style* (pp. 217-236). New York: American Elsevier.
- Biderman, A. D. (1990). The Playfair enigma: The development of the schematic representation of statistics. *Information design journal*, 6, 3-25.
- Buck, J. L., & Proulx, P. (1991). Statistics and noncognacy. *Proceedings of the seventeenth LACUS Forum 1990*, 472-480.
- Derbyshire, D. (1986). Topic continuity and OVS order in Hixkaryana. In J. Sherzer & G. Urban (Eds.), *Native South American discourse* (pp. 237-305). Berlin: Mouton de Gruyter.
- Dyen, I. (1975). Lexicostatistics so far. In *Linguistic subgrouping and lexicostatistics* (pp. 75-90). The Hague: Mouton.
- Embleton, S. M. (1986). *Statistics in historical linguistics*. Bochum: Studienverlag Dr. N. Brockmeyer.
- Frischer, B. (1991). *Shifting paradigms: New approaches to Horace's "Ars Poetica"*. Atlanta: American Philological Association.
- Lord, R. D. (1958). De Morgan and the statistical study of literary style. *Biometrika*, 45, 282.
- Mosteller, F., & Wallace, D. L. (1984). *Applied Bayesian and classical inference: The case of "The Federalist" papers*. New York: Springer-Verlag (Original work published in 1964 as *Inference and disputed authorship: "The Federalist"*).
- Skinner, B. F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology*, LIV, 64-79.
- Williams, C. B. (1956). A note on an early statistical study of literary style. *Biometrika*, 43, 248-256.
- Yule, G. U. (1968). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press. (Original work published in 1944).