COMPUTERS IN CONTENT ANALYSIS

Janet M. Vasilius

Traditionally, scholars felt "the humanities should be concerned with quality and with individual man, computers with things in quantity or men in the mass"; humanists dealt with words, scientists with numbers, and division of methodology was de rigéur.[1] Fortunately, humanists found they needed the scientific method, and the scientists discovered that numbers were meaningless without application. The increased use of computers in the humanities, coupled with the increased availability of computers to use, are a prime result of decreased isolation between academic disciplines. However, due to a history of rejection, problems in computer use persist.

Content analysis is a research technique particularly suited to the communication scholar, although content anlaysis itself is not restricted to communication. Most content analysis studies have been concerned with journalism, political affairs and psychotherapy. Good content analysis should avoid equally the "counting" phenomena which so trivializes many projects, yet keep its methods above "impressionistic" analysis by reading on and not between the lines.[2] Content analysis should also avail itself of optimum practicable research methods, like the computer.

While an occasional individual argues that the computer is but an extension of the cuckoo clock, and therefore, fairly antique, computers did not really get started until the World War II technology boom. Since the post-war period, computer technology developed rapidly. Initially, computers were designed to perform a series of arithmetic operations, and access to these procedures required much programmer sophistication. For a computer to achieve a square root, for example, it followed a series of simple operations which, though reliable, were comparatively time consuming. When STRETCH computed single-operation square roots in one-fifth the usual time, rapid reduced operations were initiated.[3] Fortran was another major breakthrough. Instead of highly detailed machine language and a professional programmer unfamiliar with individual project needs, the informed researcher could do his own programming.[4]

Although initially few in number, some behavioral scientists discovered that almost any statistical tool adapted to computer usage.[5] Eventually, prepackaged programs like Biomedical, or BMD, and the Stastical Package for the Social Sciences, or SPSS, were developed specifically for such statistical applications. More languages, like SNOBOL, and more functions were added to computer capabilities until the numbers and specialized languages of a computer could be substituted for the words and symbols of the content analyst. Since that time, with one exception, innovation has consisted of expansion of the initial techniques.

In 1963, content analysis by computer was boosted by an imaginative analysis of disputed Federalist Papers. The authorship of the papers had been unresolved by prior content analysis, but the expanded capabilities of the computer dealt with the 100,000 words and the minute factors of style as no human coder could cope.[6] As computer use expanded to data organization and reduction, hypothesis seeking and hypothesis testing, three major areas of use developed:[7] numerical information retrieval and simulation.[8] While not directly supportive of content analysis, these functions are used in, if not primarily for, such analysis; content analysis, with its limited material, forms a subset of information retrieval programs.

The year of 1966 brought the General Inquirer, a group of proceedures that form the basis for computer-assisted content analysis. Since this was the first and the last major innovation designed for content analysis, a closer examination of the procedures is warranted.

The General Inquirer maximizes the ability of a computer to compare and rearrange information, rather than merely to perform arithmetic. The program is actually a number of programs grouped under one label; program functions differ, and each program is user unique. The functions of the program can range from compilation of a concordance for editorial reference to a multi-step evaluative assertion analysis. Any program is applicable to literary analysis as easily as international affairs, provided that the program is appropriate to the research design. In addition to pre-existing programs, the General Inquirer expands every time an individual researcher creates his own program. These individual programs then are added to the General Inquirer and are available for general use. Thus, a researcher does not use _the_ General Inquirer program, but _a_ General Inquirer program.

Philip Stone describes the General Inquirer as a set of computer programs to:

> a) identify systematically, within text, instances of words and phrases that belong to categories specified by the investigator;
> b) count occurences and specified co-occurrences of these categories;
> c) print and graph tabulations;
> d) perform statistical tests; and
> e) sort and regroup sentences according to whether they contain instances of a particular category or combination of categories.[9]

The investigator must set the categories, specify the procedures, and analyze the results; the computer performs the clerical tasks. In this regard the computer must be seen as an aid to, but not a replacement for, the researcher.

General Inquirer content analysis must begin with a good research design. The data must be organized and coded so that is can be efficiently transferred to punch cards, magnetic tape or whatever method of data input is used. A clear set of coding instructions is crucial, because a computer cannot detect coding difficulties as can a human coder. Concurrently with the data preparation, the researcher must select the computer program to be used; i.e., evaluative assertion analysis, a technique used to determine various components of attitudes might be selected for use in a persuasion study.[10]

Each version of the General Inquirer system has at its core a dictionary developed to identify the tags representing the investigator's theory. One such dictionary, developed by Holsti, places words into Osgood's three dimensions--evaluation, potency and activity. A semantic differential scale is constructed, and each word is given numbers corresponding to the scale for each dimension; thus, "abandon" would be -2,-2,-3, abolish would be 2,3,2 and accomodate would be 2,1. The scale ranges from -3 to 3 and does not register 0.[11]

However, because frequency alone might be insufficient, syntax of theme codes could be required as a secondary requirement of the data. Additional program factors could include separate scores for the sentence complexities of quality and performance, an automatic score reversal if a negative is within the sentence, weighted intensity scores, statistical procedures, or a new or second dictionary may be applied for the same data. The programs are limited by available time, and the number of print-outs the investigator is willing to read.

Computer programs in the General Inquirer system have been used for projects as diverse as classifications of pottery or suicide notes. Good indices of diversity are the available dictionaries: the Harvard 3rd Psychosociological Dictionary is currently the largest, containing enough tags and categories to cover over 98% of most materials written in English; the Semantic Differential Dictionary mentioned above was developed for analysis of political documents; the Santa Fe 3rd Anthropological Dictionary allows cross-cultural comparison of folk tales; The Therapist Tactics Dictionary allows interview analysis; a "need achievement" dictionary is used for both interviews and written documents; a set of dictionaries aids analysis of products and corporate imates; a political value list exists; social class can be determined; WAI catalogs responses to "Who am I?"; folklore dictionaries deal with Icarus legends, alcohol use, Mayan jokes, Ge methology and pot. Language and cross cultural dictionaries abound, along with professional and therapeutic programs. Moreover, dictionaries are interchangable and reduplicative, as long as the theoretical assumptions are maintained.[12]

Programs currently in use for content analysis fit into the General Inquirer system, implicitly or explicitly. While the applications, programs, and dictionaries are continually updated and expanded, the Inquirer remains the major development, and probably will remain so until the computer takes over total analysis. The OCCULT program can scan texts directly; Shakespeare's intent in the first act of Hamlet can be deduced; the morality of a progression of party platforms can be determined; essay style can be classified; personal correspondence can be examined for personality traits; election results can be predicted on the basis of bias analyzed in local newspaper editorials; maps can be read, textbooks can be evaluated; and the psychotic can be diagnosed. Computers have even demonstrated an ability to "hear" voices and "see" handwriting for some time.[13] With the computer thus triumphant, what remains?

Plenty. Regardless of the progress that has been made, computer phobia and computer failings combine to preclude a total shift to the mechanical monsters.

Consider first the prime advantages of the computer: the savings in time and money.

> Given a desk calculator and a very large supply of pencils and paper, the individual researcher..could quite probably accomplish any task that a computer could. But a computer can accomplish in 60 seconds what might take an individual several days to do.[14]

In addition, the individual with the sensitivity to code well could easily become bored, or worse. When Lane Cooper prepared the Cornell Wardsworth concordance he did so by "lashing on squadrons of graduate students, discontented Ithaca housewives, and junior colleagues (incidently, three of whom died during the operation) into completion in one year."[15] However, the alternative to this type of drudgery is another type of drudgery, that of coding, punching, proofreading, defining routines, tracking materials through the process, watching for program bugs and organizing the output.[16] For every large study made feasible by the computer, there is a small study made silly by the machine. The single-shot study may not justify the expense of the keypunch operator, nor, if it is small, may the computer time be justifiable.

It is undeniable, however, that the computer makes possible projects previously unattempted. For example, "those who conducted the attribution studies on the Federalist Papers, the Letters of Junius and the Epistles of St. Paul dealt in millions of words and lived to tell about it."[17] By contrast, the tabulation difficulties of the RADIR project most possibly discouraged other non-computerized projects of such comprehensiveness.[18] Also, in addition to sheer physical size, the complexity of the data may make hand coding impracticable in terms both of time and reliability.[19] A computer can find and code items bypassed by an individual, assuming the initial data is punched properly, thus greatly increasing reliability.

But while the computer is competent at getting a lot of information from a lot of data, and a lot of information from a moderate amount of data, it is ineffective, particularly on a cost/benefit basis, at finding a little information from a lot of data. It is frustrating to both machine and analyst to sort through volumes of irrelevant material, such as a press reference to Governor Jerry Brown's superior intellect, when use of an index or sampling could better serve the function, to say nothing of the budget![20]

If the data will require different analyses, punched cards can save a great amount of time. The danger lies in the temptation to overuse the data on various "fishing expeditions." If the purpose is worthwhile, however, the cost can be minimized with successive reuses.[21] Likewise, more than one scholar can use a punched deck; thus the study can be spread over time and distance and be used by multiple investigators. The drawback is a lack of centralized information about possible data transfer and lack of clarity about the appropriateness of the data for each experimenter. A library, especially for punched literary texts, would be invaluable. Dictionaries, also, which now may be developed for a single project and then forgotten, could also be pooled.[22]

Besides data preparation, interpretation also raises questions about computer use. For problems of time and space, such as news analyses, measuring the data with a ruler may be easier, cheaper and more accurate than a sophisticated word count program for the machine. Thematic analysis is open to bias if the themes are identified and coded prior to punching,or liable to triviality if all themes are punched and processed.[23] The simpler word count and readibility processes, while less prone to coder error, have automatic limits without contextual referents; attempts to compensate can lead to endless word lists with correlations beyond a level of relevance. The leftover list, on which both mistakes and words no included in the dictionary appear, provides a valuable mechanism to check reliability and, if necessary, reformulate the dictionary if significant words are omitted.[24] However, incidence of "forgotten" words could be misleading until the print-out analysis is completed and encourages mushrooming of dictionaries.

While an inappropriate dictionary choice, or incomplete dictionary formulation can be recognized and corrected fairly easily, less obvious errors can pass unnoticed. This is particularily true if the investigator did not write his own program. The output can be totally meaningless, and may never be noticed![25] Cluster sampling may lead to overestimating significance, but reduced sample size may threaten the vaildity while, as indicated above, too large a sample may obscure results.[26] Pre-editing to control the sample is a poor procedure.[27] Editing is slow, costly, and admits experimenter bias into the data selection process. Homographs, or, multiple uses of the same word/symbol can reduce contextual interpretation to inanity; the circus bear, Wall St. "bear," pre-breakfast "bear" do not "bear" closing comparison the each other, let alone "bearing" away items, "bearing" a strain, "bearing" to the left, or "bearing" in mind homographic considerations. A disambiguation program must be added to avoid connotative error.[28]

The natural "stupidity" of the computer is a major stumbling block. The ductility of the machines Kerlinger explains, means that they are "extremely useful, obedient and reliable servants, though one must remember that they are utterly stupid."[29] If "people cannot count, at least not very high, one must remember that computers cannot think at all.[30] The computer unit of analysis is the single symbol; multiple passes and programs are needed to accomplish what a human can do in a single operation; the cards are slow and bulky and must be pre-thought or coded manually. All this places a great burden on the researcher. The computer may be reliable, but the computer cannot tell you anything about reliability; therefore, instructions must be written with utmost clarity and any confusion anticipated before the fact, both validity and reliability must be checked whenever possible; and duplication is mandatory.

However, exacting as the computer dictates may be, they are really little more rigorous than the standards the experimenter should be following anyway. Thus, the stupidity of the computer acts as a check against the laziness of the human. The precision of computer demands may initiate redefinitions of accepted theory. When Karl Kroeber told the university programmers he wished to analyze literary style, they responded with an inquiry as to what he meant by "style," as a result, Kreober has been "...trying to find out what I do mean by style...forced to recognize how little I know about my own subject...forced to criticize assumptions I had used unthinkingly for years.[31] Such re-evaluation is essential when doing any kind of content analysis.

The greatest barriers to effective computer use, however, do not come from machine flaws and requisites, but from the users. The anti-machine mentality persists, and even where it has departed, it has left residual misapprehensions. Computers are desirable because they reduce research time and, supposedly, allow more time and material access for research. However, in the first year after the General Inquirer was widely available, only 0.2% of literary scholars were conducting computer assisted research, and, of these 120 studies, all but 7 were concordances, word lists, translations, and linguistic studies.[32] Beginning researchers are attracted to the computer because machine thoroughness indicates high reliability and computational accuracy.[33] Yet it is these researchers who "rarely know anything beyond high school algebra and mostly do not know that much" and thus cannot appreciate the accuracy they demand.[34] Simiiliarily, while being attracted to sophisticated program possibilities, the novice tends to use packaged programs or relies upon professional programmers. Neither course is desirable. The professional computer programmer knows computers, but not the methodology of content analysis in the behavioral sciences. The package program may lack necessary and desirable analysis.[35]

A second type of researcher is the non-user. Boggling as he finds the computer, assurances that the SPSS or BMD programs are designed for the novice fall on deaf ears. Machines are basically incompatible with the humanistic researcher, the reasoning flows, and, in any event, a technician could be hired if needed. This individual likes to speak of truth, rather than statistics, and, if he uses content analysis at all, will do so unassisted by computer.

A third type of researcher is equally as bad, but in an opposite direction. Fast in the grips of the "Law of the Instrument" he subjects every design to computer scrutiny, regardless of applicability.[36] The RADIR study claimed

> Content analysis is specious both when used to justify a precision that is not needed and also when used to justify a position that is unusable.[37]

Others, such as Kerlinger, Holsti, Gerbner and Milic, extend the analysis to computer overuse.

Ideally, the content analyst would be a latter-day Renaissance man: skilled in research design, able to use all known statistical methods without error, filled with insight and creativity, able to program a computer unaided and endowed with wisdom, discretion, unlimited funds and a battalion of research assistants. However, such is never the case, and the individual rarely has time to master his own area, let alone computer technology. The other alternatives are equally silly: ignoring a computer will hardly make it go away, and your research will suffer in the meantime; even with the funds to hire a technician, it is no guarantee of accuracy for your problems; packaged programs may be unavailable or inappropriate and the subsequent analysis would yield little.

A balance must be struck. The researcher must first master the details of his own design. Secondly, some experience with computer programs and languages is necessary to tell others your needs as regards the computer. And, finally, humanist and scientist alike must minimize their differences and use the computer freely but appropriately, to encourage the development and dissemination of programs useful and accessible to all. The content analyst, or any researcher, has no grounds to criticize computer poetry until he has succeeded in mastering computer. The need for computer acceptance is indicated by Kerlinger.[38]

> Scholars in virtually all disciplines have no choice: they must use and master the computer. Indeed, it can even be said that the scholar of 1975 will be...obsolete if he does not understand and use the computer in his work.

Perhaps content analysis, or any other procedures will soon be interfaced and transmitted at the flick of a switch.

Until that day, efforts must be made to: 1) Improve the computer so that symbols are as easily manipulated as numbers; 2) Reduce data preparation; 3) Simplify so that a layman can more easily learn to program; 4) Expand access to data and programs. Simultaneously, the researcher must: 1) De-mythologize the computer as God or foe; 2) Learn to program the computer, or at least communicate with computer technicians; 3) Apply more creativity; and 4) Use frequently.

NOTES

Janet M. Vasilius is an Instructor in Speech and Theatre and Assistant Director of Forensics at Middle Tennessee State University.

[1] E. A. Bowles, Computers in Humanistic Research,(Englewood Cliffs: Prentice-Hall, 1967).

[2] I de S. Pool, H. Lasswel, et al., The Prestige Press, (Cambridge: M. I. T., 1970).

[3] E. A. Goldstine, Computer Newsletter, II (1965): 154.

[4] F. N. Kerlinger, Foundations of Behavioral Research,(New York: Holt, Rinehart, and Winston, 1973).

[5] Richard W. Budd, R. K. Thorp, and L. Donohew, Content Analysis of Communication (New York: MacMillan Co., 1967).

[6] F. Mosteller, and D. L. Wallace, "Inference in an authorship problem,"Journal of the American Statistics Association, (1963) p. 58.

[7] Harold Borko, Computer Applications in the Behavioral Sciences, (Englewood Cliffs: Prentice Hall, 1962).

[8] Kenneth Janda, "Some computer applications in political sciences," Computers and the Humanities, II: 12-16.

[9] Phillip J. Stone, D. C. Dunphy, et al., The General Inquirer: A Computer Approach to Content Analysis,(M. I. T.: M. I. T. Press, 1966).

________________________________ User's Manual, (M. I. T.: Press, 1966).

[10] O. R. Holsti, Content Analysis for the Social Sciences and Humanities, (Reading: Addison-Wesley, 1969).

[11] O. R. Holsti, "An adaptation of the General Inquirer for the systematic analysis of political documents," Behavioral Science, IX: 382-388.

[12] Stone ; Holsti, *Content Analysis* .. ,.

[13] E. E. David, Jr. and O. G. Selfridge, *Proceedings of the IRE*, (New York: New Press, 1962).

[14] Budd, p. 91.

[15] S. M. Parrish, "Computers and the muse of literature," *Computers and the Humanities*, II (1965): 57.

[16] Jacov Leed, *Computers and the Humanities*, I (1966): 12.

[17] Louis T. Milic, "Winger Words: varieties of computer applications in literature," *Computers and the Humanities*, II: 24-32.

[18] G. R. Petty, and W. M. Gibson, *Project OCCULT*, (New York: NYU Press, 1970).

[19] Holsti, *Content Analysis*... , p. 192.

[20] H. P. Iker, "Historical note of the use of word-frequency continuities in content analysis, *Computers and the Humanities*, VIII (1974): 93.

[21] Holsti, *Content Analysis*... , p. 152.

[22] Thomas Carney, *Content Analysis*: a Technique for Systematic Inference from Communications, (Winnipeg: University of Manitoba, 1972).

[23] Holsti, *Content Analysis*..., p. 154.

[24] D. P. Dunphy, and M. Smith, "The General Inquirer," *Behavioral Science*, X (1956): 468-480.

[25] *Kerlinger*, p. 706.

[26] P. Auld, Jr. and E. J. Murphy, "Content analysis studies of psychotherapy," *Psychological Bulletin*, LII (1955): 377-395.

C. W. Backamn, "Sampling mass media content: the use of cluster design," American Sociological Review, XXI (1956): 729-733.

[27] P. Emmert, and W. Brooks, Methods of Research in Communication, (New York: Houghton-Mifflin, 1970).

[28] Holsti, Content Analysis...p. 163.

G. Gerbner, O. Holsti, K. Krippendorf, W. Pailsley, and P. Stone (eds.), The Analysis of Communication Content, (New York: Wiley, 1969).

[29] Kerlinger, p. 710.

[30] Mosteller, p. 70.

[31] Karl Kroeber, Computers in the Humanities, VIII (1967): 37.

[32] Milic, p. 28.

[33] Budd, p. 95.

[34] G. L. Cowgill, "Computer applications in archaeology," Computers in the Humanities, II (1967): 17-24.

[35] Kerlinger, p. 707.

[36] Holsti, Content Analysis..., p. 194.

[37] Petty, p. 192.

[38] Kerlinger, p. 709.