VOLUME I | NR. 3 | OCTOBER 1968

STUDIES

MOUTON

COMPUTER STUDIES

in the Humanities and Verbal Behavior

EDITORIAL ADDRESS

Computer Studies in the Humanities and Verbal Behavior is jointly sponsored by the universities of Colorado, Kansas and North Carolina. Manuscripts: LEWIS SAWIN, Department of English, University of Colorado, Boulder, Colorado, U.S.A. General correspondence: FLOYD R. HOROWITZ, Department of English, University of Kansas, Lawrence, Kansas, U.S.A.

Computer Studies in the Humanities and Verbal Behavior appears quarterly in issues of approximately sixty pages. Four issues constitute a volume. The subscription price is $ 10.00/f 36.00 per year per volume; single issues cost $ 3.00/f 10.50. Back issues as well as subscriptions and single issues can be ordered from every bookseller or subscription agency, or directly from Mouton & Co., P.O. Box 1132, The Hague, The Netherlands.

DATA-TEXT: A Simple and Flexible Programming System for Historians, Linguists, and Other Social Scientists

JUDITH E. SELVIDGE and THEODORE K. RABB
Harvard University Computer Center
Department of History, Princeton University

The purpose of the paper is to make a description of the DATA-TEXT System widely accessible to historians and linguists who are sensitive to the ability of computers to store, retrieve, and analyze large bodies of material previously too unwieldy to handle.

An historical study example is used to illustrate the DATA-TEXT program. Specifically, the problem examined is a study of the ecclesiastical princes of the Holy Roman Empire between 1555 and 1612. Data on social background, education, geographic movements, career, and activity while in office of every bishop and abbot who held a landed see or abbey in this period was gathered, reduced to basic categories, coded numerically and then subjected to several analyses by the DATA-TEXT. A few of these analyses are discussed in detail and used to illustrate the more elegant and advantageous features of the program. These features are: ability to modify values of items in the study, recode instructions, one-way analysis of variance, included mathematical functions, cross-tabulation, options for punch-card output and tape input and output, frequency analysis of variables, and provision for updating periodically a tape masterfile.

Finally, the program's use in textual editing and analysis of material presented in the form of strings of words is very briefly discussed.

The most obvious assistance that computers can give to historians and linguists is to provide them with the ability to retrieve, store, and analyze large bodies of material which previously have been too unwieldy to handle. There are huge concentrations of historical data — election returns, proceedings of law courts, records illustrating the careers of members of important institutions, trade figures, and many others — which have always been far too cumbersome to investigate *in toto*. An essential part of any such study has to be a numerical analysis; but, even though this approach enables one to bring to the surface a new level of hitherto unused evidence, one cannot organize the information, let alone make accurate calculations, without modern data-processing machines.[1] The same is true of that branch of linguistics concerned with detailed textual analyses. The first steps — taking word counts and making the simplest arrangements of word groupings — are tedious and laborious tasks, in which precision becomes particularly elusive when lengthy texts are involved. Here again the computer offers qualities of speed, flexibility and accuracy which render previously impossible projects feasible.

Unfortunately, the mere knowledge that computers have an almost limitless capacity for information retrieval and evaluation is not enough to solve the historian's and the linguist's problems. The principal use to which computers have been put in the early years of their existence has been mathematical, and the principal users have been scientists and mathematicians. Consequently, historians and linguists find that there are very few standard programs available to help them in their investigations. For this reason it seemed advisable to make a description of the DATA-TEXT System widely accessible. As will be seen, it can be programmed very simply to perform most of the cross-tabulations, correlations, and simple statistical analyses which are essential to any historical study dealing with a large mass of data; and it can also perform the basic functions of linguistic analysis. This paper will concentrate on the advantages the system holds for the historian — and in fact for all social scientists: the example that will be used happens to be a historical study, but its basic principles could be employed in sociology, economics, psychology, educational theory and a number of related

[1] The opportunities opened up for historians by computers and some of the difficulties that remain are discussed at some length in Theodore K. Rabb, *Enterprise and Empire: Merchant and Gentry Investment in the Expansion of England, 1575-1630* (Cambridge, Massachusetts, 1967), especially the Introduction and Chapter 3.

fields — but there will also be a short section explaining DATA-TEXT's suitability for the linguist.

The fundamental requirement of the historian is for cross-tabulations and correlations between multiple variables. The categories of information he can retrieve are often strictly limited. If the basic unit is a person, the researcher may be able to specify no more than some of the chief dates of the individual's life, his social background, education, geographic movements, financial resources and expenses, and public career. Within each of these categories the alternatives may be legion (one could, for example, code or assign a number to every high school and college in the United States, if such a distinction were meaningful), but the categories themselves are usually few in number, particularly if the subject is taken from a period more than 200 years ago, when the documentation is likely to be sparse at best. One may have a few pieces of information about a large number of men (or court cases, or trading voyages), but detailed material about each of the units is unlikely to have survived.

The most useful analysis of this type of data is cross-tabulation of the crucial variables. Did the social background determine the career pattern? Or the education? Or the life span? Were profits more likely to accrue from a trading voyage with a high original investment or a low investment? In what periods were a court's sentences more severe and when less severe? All such questions are best answered by cross-tabulations and simple correlations. It is rare and almost invariably in recent periods that the historian has sufficient data to perform more intricate analyses. At most he will want to apply a few basic statistical tests (such as the calculation of chi-squares) to the tables that he has compiled. Though the computer will be doing little more than a mammoth job of sorting for him, it will allow him to assemble, organize, and evaluate evidence which otherwise would be inaccessible.

As will be seen, DATA-TEXT performs all these functions, and yet requires a minimum of effort and programming expertise from the user. Many of the statistical tests are provided automatically, and both the description of the data and the creation of the tables can be accomplished in a few easy steps.

The DATA-TEXT System[2] is a large-scale computer program primarily designed to analyze data obtained from research in the humanities and social sciences. The author of the System, Dr. Arthur S. Couch of the Social Relations Department of Harvard University, envisages the reduction of data by a high-speed electronic computer as a two-stage process. The first stage of the computer's work is the modification, or editing, of the information collected. Statistical analysis of the edited data is the second stage. The DATA-TEXT System facilitates these two tasks by anticipating the kinds of editing and analysis to be done without involving the researcher in the computational details. The name "DATA-TEXT" signifies that the System has provisions for analyzing both numerical and textual information, but most of the description which follows will concentrate on its more widely used capability — dealing with numerical data. The methods of implementing the System are fully described in the DATA-TEXT manual, which can be understood without previous programming knowledge; an experienced programmer, however, is at an advantage and can become conversant with the operation of the System in a matter of hours.

The editing capacities of DATA-TEXT are extensive, and the great flexibility of the System is due mainly to its ability to take into account the second thoughts of the researcher. When large amounts of data are gathered, some items may turn out not to be recorded in exactly the form best suited to the needs of computer analysis. Furthermore, errors are frequently introduced as material is transcribed onto Hollerith punched cards or magnetic tape. To verify or correct the data by hand or by using specially written programs can be an extremely tedious and expensive job. The DATA-TEXT System turns most of this work over to the computer. The editing options that are available include the creation of variables defined as mathematical functions of the original coded information; the deciphering of card columns containing multiple punches; the recoding of alphabetic information into numerical form; and a check to insure that the values of a variable fall within some pre-determined range. Monitoring the range of values given to a variable is one of several techniques which can be used to locate inconsistencies in the data. The researcher can specify that certain inconsistencies are to be considered erroneous. Should any such errors occur the program will identify the unacceptable observations and will omit them from any further calculations.

Once the data have been edited the program can be instructed to continue and to perform an analysis; or else the computing may be interrupted to allow for the correction of errors in the input data. Among the analytic techniques supplied by the System are correlation among variables, calculation of means, standard deviations, third and fourth moments, many-dimensional cross-tabulation, contingency tables, frequency analysis, one-way analysis

[2] The development of the DATA-TEXT System has been supported by grants from the National Science Foundation (GS-1424) and the Laboratory for Social Relations, Harvard University.

of variance, t-test and factor analysis. One or several of these analyses may be requested and the ease with which they can be initiated is one of DATA-TEXT's most attractive characteristics.

The best way to demonstrate the uses of the DATA-TEXT System is to examine a typical problem to which it was recently applied. This was a study of the ecclesiastical princes of the Holy Roman Empire between 1555 and 1612, undertaken by Miss Sarah E. Gibbard, a graduate student at Harvard University, for a Seminar taught by Professor Rabb.[3] She gathered data on the social background, education, geographic movements, career, and activity while in office of every bishop and abbot who held a landed see or abbey in this period. This information was reduced to basic clear-cut categories and coded numerically. The study was restricted to a short time span because the project had to be completed in three months, but it could be enlarged to include all the bishops and abbots from the year 1000 to the present. The purpose of the investigation was to gain familiarity with DATA-TEXT's usefulness in making cross-tabulations, and to answer questions like the correlation between social class and age at elevation. What follows are descriptions of a few of the analyses that were performed.

Control Cards

```
*DECK BISHOPRIC STUDY FOR MISS GIBBARD
  DEC 66
*FORMAT (A4,2(F1.0,F2.0), 4F1.0,
  2F2.0,6F1.0)/UNIT, DATA (16)/SEQ(1)
*VAR(1) = X(1) = BPRIC LEV
*VAR(2) = X(2) = BPRIC LOC
*VAR(3) = X(3) = FAMILY CLASS
*VAR(4) = X(4) = FAMILY LOC
*VAR(5) = X(5) = PREDECESSOR
*VAR(6) = X(6) = SUCCESSOR
*VAR(7) = X(7) = JESUIT ED
*VAR(8) = X(8) = GERMAN UNI
*VAR(9) = X(9) = YR OF BIRTH
*VAR(10) = X(10) = YR ELECTED
*VAR(11-16) = X(11-16) = DECADE
*VAR(17) = RECODE X(10), CODE (A) =
  YR ELECTED
*CODE(A) = (BLANK = ERROR)
*PRINT UNIT IF VAR(9) GREATER THAN
  VAR(10)
*COMPUTE FREQUENCIES(1-9,11-16,17)
```

The statement beginning "*DECK" simply provides a title which will appear on all DATA-TEXT output. The part of this list of instructions most difficult to explain is the next instruction, the FORMAT statement. Its purpose is to describe to the computer exactly the layout of the input data on the punched data cards or magnetic tape. The means employed by DATA-TEXT for this explanation is very similar to the method of the programming language, FORTRAN. Format statements are perhaps the most unnatural part of the FORTRAN language, which otherwise imitates ordinary English fairly successfully. The meanings of various parts of the *FORMAT statements are explained at length in the DATA-TEXT manual; it will suffice in this example to say that the first part of the statement gives the location by card column of the characters which comprise the input data (items 1 through 16), and that the first four columns of each card contain a "unit" or "observations" number which uniquely identifies that card. As the various items of one observation, or "data case", are read in, the program will automatically store these values in an array "X", and the items must henceforward be referred to as "X(so-and-so)": the first item of data becomes X(1), the second becomes X(2), and so on up to, in this example, X(16), the last item on each card of input data.

The DATA-TEXT System assumes that we will want to edit or modify all the items in the study. The modified values will then be stored in an array named VAR, and all analyses will be carried out on the VAR quantities. Consequently, each X item which is to be included in the study must be entered into the VAR array even if the value is not modified. In the example above, the statement *VAR(1) = X(1) = BPRIC LEV serves to store the X(1) value in the VAR(1) part of the VAR array. The alphabetic characters "BPRIC LEV" appearing after the second equals sign will also be saved and will be printed out whenever VAR(1) is mentioned in the later analysis. This optional feature allows us to assign labels to all the variables, making the printed output much easier to read. Up to 24 characters may be used for each label. The abbreviated title here stands for BISHOPRIC LEVEL. (These were divided into Archbishopric (1), Bishopric (2), and Abbacy (3).) The next nine statements are identical in function to the one just discussed with only the item numbers and the labels changed. The meanings of most of the variables should be clear from their abbreviations: the second variable, for example, is a code number which designates the location of the Bishopric. (Each of the ten imperial "circles" — administrative units — was given a number.) The same code is used in VAR(4) to classify the location of the subject's family estate. The appointment to a Bishopric was often kept in the family. Variables (5) and (6) take this into account by recording the relation-

[3] The authors are grateful to Miss Gibbard for permission to cite a part of the program she worked out in conjunction with Miss Selvidge as the main example in this paper. The funds for her project were provided by a grant form the Department of History at Harvard University.

ship, if any, of the subject to his predecessor and successor. His education (e.g., which Jesuit college, or which university) is described by the values of the next two variables. It was also decided to compare the composition of the group at the beginning of each decade, from 1560 to 1610. To facilitate the partitioning of the group into various sub-groups, or "decade cross-sections", for this analysis six variables, VAR(11-16), were set up to correspond to these six dates. Each subject was given the value "one" or "zero" for each of the six variables according to whether he was or was not in office at that date. It should be noted that all X values are entered into the VAR array with the single statement *VAR(11-16) = X(11-16) = DECADE. This condensation reduces the number of control cards to be written. All the variables (11-16) will have the same label, "DECADE". However, the System will assign numbers in sequence to this label so that VAR(11), for example, will have the title "DECADE NO. 1", and VAR (16) the title "DECADE NO. 6".[4]

A recode instruction is introduced in the specification of VAR(17). It was decided that the datum "year of election to bishopric" was so important that any subjects for whom this information was missing should be eliminated from the study. The statement, *VAR(17) = RECODE X(10), CODE(A) = YR ELECTED, where *CODE(A) = (BLANK = ERROR) takes the value of X(10), the year of election, and recodes it, setting blanks (missing information) equal to an ERROR. All non-blank values will remain unchanged. This recoded X(10) is then stored in the location — VAR(17). During the execution of the program, for each unit, or "observation", all the VAR values will be checked to see if an error condition exists. When an error is found, the number of the unit with the erroneous observation is printed out along with the message "ERROR VALUE FOR VAR(so-and-so)". In this program, an error specification having been set up only for VAR(17), no error will be detected unless VAR(17) is blank.

Certain checks of the data can even be made without using a RECODE statement. For example, the instruction:

```
*PRINT UNIT IF VAR(9) GREATER THAN
  VAR(10)
```

calls for the printing out of the unit number if, referring back to the VAR specifications, the year of birth is larger, that is to say later, than the year of election to the Bishopric, implying either an error in the data or prenatal elevation. A conditional or "IF", statement can cite, besides the "GREATER THAN" relationship mentioned here, the operators "LESS THAN" and "EQUAL". Compound conditionals are also permitted using the inclusive operator "AND" and the exclusive "OR".

We may want not only to check the accuracy of the input data but also to create new variables which are arithmetic or logical functions of the original items. Three new quantities which may be obtained using the data in our example are:

```
*VAR(18) = X(10) - X(9) = AGE AT
  ELECTION
*VAR(19) = SUM(X(11-16)) = NO. OF DEC
  IN OFFICE
*VAR(20) = RECODE X(3), CODE (B) = NEW
  CLASS DESIGNATIONS
*CODE(B) = (1,2=1/3=2/4,5=3)
```

The AGE AT ELECTION is set equal to the year elected minus the year of birth. Addition, multiplication, and division as well as subtraction and combinations of all these may be used to describe variables. Furthermore, certain mathematical functions are provided by the program such as sum (see VAR(19)), square root, absolute value, arcsine, and logarithm.

A value of "one" for any of the variables 11 through 16 indicates that the subject was in office during that specific year. To find the total number of "decade cross-section" entries in which each bishop appeared we calculate the sum of X(11) through X(16).

The statement which specifies VAR(20) is another example of a RECODE instruction. Suppose that "family class", VAR(3), is indicated by a number which has been given values from 1 to 5. where 1 signifies a member of the Hapsburg family, 2 a member of a princely family, and so on down to 5 which designates a non-land-owning peasant. If we decide that for some analyses this breakdown by class is too fine a grouping, we may then re-group the subjects into three categories, placing members of the old group 1 and 2 into the new group 1, 3 into new class 2, and 4 and 5 into the new group 3. This reassignment is effected by the two statements:

```
*VAR(20) = RECODE X(3), CODE (B) = NEW
  CLASS DESIGNATIONS
*CODE(B)=(1,2=1/3=2/4,5=3)
```

[4] Listed here are only a few of the categories of information that were assembled about each bishop and abbot. The following were some of the other data punched onto the card: position in the family (i.e., heir, younger son, bastard, etc.); number of members of one family who held the same office successively; religious affiliation of tutors; other offices held; the chief influence ensuring the election (e.g., family cardinalates); marriage; avowed religious affiliation; and Counter-Reformation activity while in office (e.g., introduction of Jesuits or Tridentine decrees, measures against Protestants, reform of abuses).

```
GIBBARD BISHOPRIC STUDY DEC 66

                                FREQUENCY DISTRIBUTIONS

VAR 1
BRPIC LEV N= 153 MEAN= 1.869  MEDIAN= 2.000 SD= 0.467 SKEW= =-0.421 KURTOSIS= 0.994
VALUE          1.00     2.00     3.00
FREQUENCY       (28)    (117)     (8)
PERCENTAGE    18.30    76.47     5.23
CUMULATIVE   0.1830   0.9477   1.0000

VAR 2
BPRIC LOC N= 153 MEAN= 5.473   MEDIAN= 6.000 SD= 3.064 SKEW= -0.163 KURTOSIS= -1.240
VALUE          1.00     2.00     4.00     5.00     6.00     7.00     8.00     9.00    10.00
FREQUENCY       (24)     (12)     (21)     (16)     (16)      (5)     (18)     (30)     (11)
PERCENTAGE    15.69     7.84    13.73    10.46    10.46     3.27    11.76    19.61     7.19
CUMULATIVE   0.1569   0.2353   0.3725   0.4771   0.5817   0.6144   0.7320   0.9281   1.0000
```

Fig. 1

Once the data have been checked and new variables created, the DATA-TEXT System performs the analyses requested. Usually one or two statements are sufficient to initiate an entire analysis. In example 1 we say:

```
*COMPUTE FREQUENCIES (1-9,11-16,17)
```

This instruction calls for a frequency analysis of variables 1 through 9, 11 through 16, and 17. It should be noted that the numbers in brackets refer to VAR assignments, and *not* to the original X items. For each of the variables in the COMPUTE FREQUENCIES command, a table will be produced giving the variable number, all its observed values in this study, and the frequency of occurrence of each of the observed values. Associated with each cell frequency in the printed output will be the percentage of the total number of observations which that cell frequency represents. (See Figure 1.)

During the execution of the program, as items from each observation are being read into the VAR array, certain quantities useful for later calculations are accumulated, namely the mean, number of non-blank entries, variance, standard deviation, skewness, and kurtosis for each variable. These values, called the "Basic Data Statistics" will be automatically printed out before the calculation of the *COMPUTE commands. In this example we would doubtless be interested in the means and variances of several variables, in particular VAR(18), the AGE AT ELECTION, and VAR(19), the NUMBER OF DECADES IN OFFICE.

The simple addition of an instruction or two will produce other statistical calculations. The correlation coefficients between certain pairs of variables may be of interest. To investigate the relationships, for example, of family class, age at election, year of election, and number of decades in office, we write:

```
*COMPUTE CORRELATIONS (3,18,10,19).
```

Simple one-way analysis of variance is another technique which is readily available to the DATA-TEXT user. For this calculation the units must be grouped according to some criterion. The F-ratio will then be found to compare the between-groups to within-groups variation from the mean for some variable. If, for example, we want to examine the difference in family class, VAR(3), from one Bishopric location to another, and the locations are coded as 1 to 10 according to the "circles" into which the Empire was divided, the statements to group the units would be the following:

```
*GROUP = 1                IF VAR(2) = 1
*GROUP = 2                IF VAR(2) = 2
*GROUP = 3                IF VAR(2) = 3
*GROUP = 4                IF VAR(2) = 4
*GROUP = 5                IF VAR(2) = 5
*GROUP = 6                IF VAR(2) = 6
*GROUP = 7                IF VAR(2) = 7
*GROUP = 8                IF VAR(2) = 8
*GROUP = 9                IF VAR(2) = 9
*GROUP = 10               IF VAR(2) = 10
*COMPUTE F-TESTS (3), GROUPS (1-10)
```

Under the same grouping a second analysis of variance on VAR(18), the age at election, can be obtained by changing the *COMPUTE card to read:

```
*COMPUTE F-TESTS (3,18), GROUPS (1-10)
```

A consolidated grouping of the units can be specified by changing the *GROUP cards. A three-way partition of the Empire (North, South and East) might be obtained in this way:

```
*GROUP = 1 IF VAR(2) = 1
  OR IF VAR(2) = 2 OR IF VAR(2) = 3
*GROUP = 2 IF VAR(2) = 4
  OR IF VAR(2) = 5 OR IF VAR(2) = 6
  OR IF VAR(2) = 7
*GROUP = 3 IF VAR(2) = 8
  OR IF VAR(2) = 9 OR IF VAR(2) = 10
*COMPUTE F-TESTS (3,18), GROUPS (1-3)
```

These groups could also have been created by the command:

```
*GROUP = 1 IF VAR(2) GREATER THAN 0 AND
  IF VAR (2) LESS THAN 4,
```

and so on for the rest of the groups.

The cross-tabulation feature of DATA-TEXT is extremely elegant and useful. It gives us a method of looking at the joint frequency of two or more variables. In the data considered here, for example, we might decide to examine the frequency of all possible pairs of values of VAR(3) and VAR(18). This would give a table in which family class was compared to age at election and would allow us to see whether or not the youngest bishops tended to come from the upper classes. The command:

```
*COMPUTE CROSSTABS (3 BY 18)
```

would produce the cross-tabulation in the form of a neat table complete with row and column titles, cell frequencies and percentages, and marginal frequencies and percentages. In addition we can request certain statistical measures of significance (such as chi-square) and measures of association like Kendalls Tau B. A variable which is to be included in a cross-tabulation must take on only consecutive integer values from 1 to N and each of the N Categories must be given a title. A variable which does not satisfy this condition must be recoded. The recode and the naming of the categories can be accomplished using a single command, the ORDER instruction. In the case of VAR(18), AGE AT ELECTION, we form a new variable:

```
*VAR(21)=ORDER VAR(18)=AGE GROUP(UNDER
  20=UNDER TWENTY/20-30=TWEN TO
  THIR/31-40=THIR TO FOR/41-50=FOR TO
  FIF/51-60=FIF TO SIX/OVER 60=OVER
  SIXTY)
```

In this way the values of VAR(18) are grouped into six categories (with the labels given above) which follow the equals sign in each of the specifications listed above. The VAR to be entered in the cross-tabulation is now VAR(21) so the instruction becomes:

```
*COMPUTE CROSSTABS (3 BY 21)
```

Category labels must also be assigned to VAR(3), though in this case the variable values are already only consecutive integers from 1 to 5. The labels are given to each group by modifying the *VAR(3) card to read:

```
*VAR(3) = X(3) = FAMILY CLASS
  (ROYAL/UPPER/MIDDLE/LOWER/PEASANT)
```

The entire set of DATA-TEXT cards would now read:

```
*DECK BISHOPRIC STUDY FOR MISS GIBBARD
  DEC 66
*FORMAT (A4,2(F1.0,F2.0), 4F1.0, 2F2.0,
  6F1.0)/UNIT, DATA (16)/SEQ(1)
*VAR(1) = X(1) = BPRIC LEV
*VAR(2) = X(2) = BPRIC LOC
*VAR(3) = X(3) = FAMILY CLASS
  (ROYAL/UPPER/MIDDLE/LOWER/PEASANT)
*VAR(4) = X(4) = FAMILY LOC
*VAR(5) = X(5) = PREDECESSOR
*VAR(6) = X(6) = SUCCESSOR
*VAR(7) = X(7) = JESUIT ED
*VAR(8) = X(8) = GERMAN UNI
*VAR(9) = X(9) = YR OF BIRTH
*VAR(10) = X(10) = YR ELECTED
*VAR(11-16) = X(11-16) = DECADE
*VAR(17) = RECODE X(10), CODE (A) = YR
  ELECTED
*VAR(18) = X(10) = X(9) = AGE AT
  ELECTION
*VAR(19) = SUM(X(11-16)) = NO. OF DEC
  OFFICE IN
*VAR(20) = RECODE X(3), CODE (B) = NEW
  CLASS DESIGNATIONS
*VAR(21) = ORDER VAR(18) = AGE GROUP
  (UNDER 20=UNDER TWENTY/20-30=TWEN TO
  THIR/31-40=THIR TO FOR/41-50=FOR TO
  FIF/51-60=FIF TO SIX/OVER 60 = OVER
  SIXTY)
*CODE (A) = (BLANK = ERROR)
*CODE (B) = (1,2=1/3=2/4,5=3)
*COMPUTE FREQUENCIES(1-9,11-16,17)
*COMPUTE CROSSTABS (3 BY 21)
```

The cross-tabulation table produced by this set of instructions is shown in Figure 2.

A number of the ingenious devices incorporated into the DATA-TEXT System will be appreciated particularly by readers who have had experience in dealing with large amounts of data on the computer. For example, when a single unit consists of several cards with different formats, the program will select the correct X values from each card even when the cards are out of order or when some cards are missing. The input routine of DATA-TEXT distinguishes between zeros and blanks, and, unless otherwise instructed in a RECODE statement, will treat blanks as missing observations. Several methods exist for treating subsections of the data. Besides the *GROUP instruction mentioned in the description of the Analysis of Variance, consideration of subsets of the data may be programmed

GIBBARD BISHOPRIC STUDY DEC 66

CELL PERCENT BASED ON COLUMN SUM CONTINGENCY TABLE NO. 1

VAR 21 AGE GROUP

VAR 3 FAMILY CLASS	UNDER TWENTY	TWEN TO THIR	THIR TO FOR	FOR TO FIF	FIF TO SIX	OVER SIXTY	TOTAL	PERCENT
ROYAL	100.0		9.5	4.9	1.9	7.1		
	2		2	3	1	1	9	5.9
UPPER		100.0	85.7	93.4	67.3	42.8		
		3	18	57	35	6	119	77.7
MIDDLE				1.6	19.2	35.7		
				1	10	5	16	10.5
LOWER			4.8		7.7	14.3		
			1		4	2	7	4.6
PEASANT					3.8			
					2		2	1.3
TOTAL	2	3	21	61	52	14	153	
Percent	1.3	2.0	13.7	39.9	34.0	9.2		100.0

Fig. 2

using a *SELECT UNIT command. If in the study of the bishops and abbots we wanted to repeat all the analyses *only* for those subjects who were in office in 1580, we would merely insert the command:

```
*SELECT UNIT IF VAR(13) = 1
```

Any units for which this value is not 1 would be omitted from the calculations.

Punched card output can be obtained simply by specifying a format and by supplying a list of the numbers of the variables to be included on the punched card. This allows us, for example, to produce new cards containing newly created variables plus those items which have been checked for consistency (ERROR values will appear on the cards as *blank* fields).

DATA-TEXT offers a great many options which enable the researcher to work with magnetic tape input and output rather than punched cards and printed sheets. In particular there is a feature which handles the problems of periodic up-dating of a master file. DATA-TEXT will deal with taped information in both binary and BCD form, blocked and un-blocked.

Text editing and analysis is performed by DATA-TEXT on material presented in the form of strings of words. An alphabetical listing of word counts and percentages can be requested for the entire vocabulary of the text being analyzed. In addition, the researcher can define his own set of "concepts" by stating which words he wishes to consider

as synonyms in order to perform frequency or other analyses. For example the cards:

```
*CONCEPTS
MOTHER = MOM, MA, MAMMA, MOTHER, MOMMY,
  OLD-LADY
FAMILY = MOTHER, FATHER, SON, DAUGHTER
*CONCEPTS END
```

will produce the two new categories MOTHER and FAMILY. "Concepts" can also be defined using the stems of words to signify the inclusion in one group of all words beginning with those letters, such as ACHIEV-, INTELLI-, etc.

DATA-TEXT was written in FORTRAN II and FAP for operation on an IBM 7094 under the standard FORTRAN Monitor System at the Harvard Computing Center. The program has recently been modified so that it will run under the IBSYS I/O System and can be used on "Direct-Couple" Machines. Other Universities have expressed an interest in DATA-TEXT, and versions of the System are currently in operation at the University of Chicago, Columbia University, and Stanford University, as well as at two or three commercial computing service organizations in New York City, the Department of Health, Education, and Welfare, and the Stanford Research Institute. Copies of the program have been requested by the University of California at Berkeley, Los Angeles, Riverside and Santa Barbara, by the Universities of Maine, Washington, Toronto, Wisconsin, Pennsylvania, and by Princeton, Cornell, Brandeis, Dartmouth and Yale. Plans are being made for the conversion of DATA-TEXT to operate on second generation computers such as the IBM 360 series. Enquiries for technical information and DATA-TEXT materials should be directed to the System's author, Dr. Couch.[5]

The DATA-TEXT System, although outstanding in many ways, is certainly not fault free. But its drawbacks are not the result of any difficulty in learning to write control cards; one quickly becomes accustomed to these rules. Problems usually arise either because of errors in the System that stem from its newness, or because of the addictive quality of DATA-TEXT, which sometimes leads the researcher to write an extremely complicated set of DATA-TEXT cards when a simple FORTRAN program would do the job equally well.[6] The first difficulty will become less significant as the program's bugs are discovered and eliminated. The second is perhaps as much of a compliment to the DATA-TEXT System as it is a criticism. Only the researcher's good judgment can save him the time wasted in applying DATA-TEXT to those few problems for which it is not suited.

[5] Dr. Arthur S. Couch, Department of Social Relations, William James Hall, Harvard University, Cambridge, Massachusetts, 02138. The post-paid price of the DATA-TEXT Manual is $ 4.00.

[6] Certain types of editing between units rather than within units are difficult to handle with DATA-TEXT. Some simple analyses, when hundreds of variables are involved, take much longer in terms of time on the computer under the DATA-TEXT System than when programmed separately in FORTRAN.

Computer Analysis of Dyadic Interaction*

GEORGE PSATHAS
Boston University

Based upon the General Inquirer system for content analysis, this paper presents an analysis of therapist-patient interaction as an illustration of various problems involved in any form of verbal interaction. The computer is useful for examining communication in a therapy interview situation. To classify the content of interaction, multiple dictionaries must be used. Three different dictionaries are described: the Interpersonal Identification Dictionary, Therapist Tactics Dictionary, and the Psychological Content Dictionary. Then the author describes a procedure to be followed in the analysis: (1) identity of speaker is first made before appropriate dictionary(ies) can be searched; (2) the sentence can be classified according to aspects of interaction strategy or tactics; (3) a general dictionary can be used to classify content; (4) tagging may be conditional on words and/or tags in the same statement or in preceding and/or succeeding statements; (5) by developing groupings of tags and sub-classifications, contingencies and inter-relationships between content categories can be summarized and calculated; and (6) sequences of interaction can be examined to plot a matrix of pro-action and reaction (a kind of "interaction map") to determine whether certain content patterns go together. In his conclusion, the author indicates that changes in the General Inquirer which facilitate this kind of analysis have already been achieved and that extensions into the analysis of in-process interaction can be anticipated.

Verbal interaction between two persons differs from written text in ways which require that its specific characteristics be considered before a computer system of content analysis is utilized.[1] In this paper, the analysis of interaction between therapist and patient in the therapy interview will be used to illustrate the problems involved and suggest solutions within the framework of the General Inquirer (GI) system for content analysis. The presentation assumes familiarity with the General Inquirer system.[2]

In discussing two person interaction, the particular kind of situation that will be analyzed will be one in which each person has a special role to play. For example, an interviewer and a respondent, an experimenter and a subject, a teacher and a student, a doctor and a patient. In such situations, each participant is involved in the performance of his own role and the support of the other's role. That is, he is aware that some aspects of his behavior are concerned with maintaining his role in the eyes of the other so that the other will regard him as the special person-in-role that he is. He also does the same for the other, that is, accords to him the recognition that the other is a special person-in-role and acknowledges that he knows that the other knows the same about him. When both parties interact in such a way, they can be said to be mutually aware of one another's role.

Before role performance can be achieved, some basic underlying conditions have to be met. The two persons

* This paper is a substantial revision of a paper originally presented at the National Conference on Content Analysis, Annenberg School of Communications, Philadelphia, Pennsylvania, 1967. The research has been supported by a grant from the National Institute of Mental Health, MH 12889. Dennis J. Arp, who has served as Project Director, and J. Philip Miller, who has planned the programming of the 360 version of the General Inquirer, have contributed generously their ideas and suggestions at every phase of this research. The dictionaries described herein have been developed by Dennis J. Arp and, under his direction, Diana Reed, Sandra Gold, Robert Miller, Edith Erickson and Joel Achtenberg.

[1] The possibility that the analysis of interaction will contribute to the solution of many of the problems of content analysis has been noted by Hays, who states "the analysis of content will achieve its greatest successes by operating with a model of the conversationalist". How conversation is generated and understood by those participating in it needs to be better understood before substantial progress can be made on this task. D. Hays, "Linguistic Foundations for a Theory of Content Analysis", paper presented to the National Conference on Content Analysis, Annenberg School of Communications, Philadelphia, Pennsylvania, 1967.

[2] The General Inquirer system has been described extensively in Philip J. Stone, *et al.*, *The General Inquirer: A Computer Approach to Content Analysis* (Cambridge, The M. I. T. Press, 1966).

are physically and temporally co-present, i.e., they share the same time and space. Further, they are mutually aware of one another's presence. This means that each is aware of the other and aware that the other is aware of him. This mutual awareness is continual, i.e., it goes on recurrently throughout the interaction. Explicit recognition is made of such continuity by the use of various indicators to show that they are attentive to one another. Because they are within close sensory range of one another, they are able to use a variety of senses. They can, for example, use eye contact or non-verbal gestures to indicate continued awareness as well as verbal indicators.

Their verbal interaction follows some patterning in regard to sequencing though it is not absolutely essential that an alternation of talk, where one talks and is followed by the other, occur. It is possible for one person to talk extensively while the other interacts only to indicate continued attention and awareness. Therefore, it is necessary to distinguish interaction from verbal interaction. The former can occur without the latter but not the converse. Vocalizations may be the channel used but any channel, vocal or non-vocal, can be represented by language symbols. For example, non-verbal gestures can be described in words or their meanings coded using symbolic notation.

The interaction we refer to is symbolic using the medium of language and shared cultural meanings for the non-verbal gestures. We are specifically concerned with the words generated by the participants though it is possible to include other aspects of the interaction by using words or other symbols to describe them. We expect that they use a shared language system in interacting with each other, though the sharing of meanings is not necessarily perfect for each word or statement. We assume only that there is *some* sharing, otherwise continuing interaction would be impossible.

Each participant intends to convey some meaning when he speaks but we cannot assume that his intended meaning is identical to the interpretation made by the other participant. We take as essential only that there is a giving and a receiving of meanings.

The giver, as he talks, is communicating not only with the other but with himself. He monitors his own statements and reflects on them such that he attempts to determine whether he is saying what he intends. He is also involved in monitoring the responses of the other and this aids him in determining whether his intended meanings are being similarly interpreted by the other. Feedback *to* the self *from* both the self and from the other with reference to the meanings of one's own behaviors is a continual process in interaction.

The use of an already developed shared symbolic system, such as language, facilitates the interaction considerably. It means that the two participants do not have to build a language. By using a language they both know, they can show each other very quickly that they understand each other. The language system has its own structure (syntax) and set of meanings (semantics). If we, as observers, share the same language, it is possible for us to adopt their same framework for interpreting meanings. We can proceed as though we "know" what they are talking about. However, in practice, this assumption is not always borne out. Our participants can be speaking in a special code, they can be developing a set of new meanings using the same words that we all know, or they can be acting as though they are communicating with one another when, in fact, they are having serious difficulty understanding each other. As observers, we cannot be completely certain of what is going on in their interaction. We need to be aware that we may jump to the conclusion that we understand what they are talking about merely because we speak the same language as they do and can assign meanings to what they say.

Determining the special idiosyncratic meanings which each participant holds is a formidable problem. We recognize that making such determinations is an important task and one which is of great concern to the therapist, for example, in trying to understand the patient. We do not propose to undertake the solution of this problem nor suggest how it might be solved in this paper. Rather, we will start at the level of shared language and proceed to operate in terms of the understandings that we have about the same language that our participants are using. We will use all that we know about the situation to assist us in understanding their interaction. If, for example, they perform special roles and if certain aspects of their behavior can be understood by reference to the characteristics of these roles as they are defined in the culture, then we will want to include this information in any interpretive schema we apply to the project of "making sense" of their interaction. The fact that they are a therapist and a patient, for example, provides us with considerable knowledge concerning some of the meanings which may enter into their interaction. The therapist is involved in accomplishing certain goals, a major one being the production of talk by the patient. At the same time, he is trying to understand the patient, discover the meanings of events as the patient perceives them, instruct and even subtly guide the patient in interpreting and understanding the meanings of these same events from a new perspective, provide interpretations of his own, suggest the connections between events which the patient may not understand, etc. He is simultaneously involved in a number of tasks

while talking about persons, events and the present situation with the patient. The content of their interaction (the *what*) is one of the concerns of both parties; another concern is their method or strategy (the *how*). We will have to give some attention to this distinction in order to understand interaction since all meaning does not reside in the content of what is being said.

We wish to work with a computer system for analyzing interaction despite the fact that we are aware of its limitations in dealing with complex data of the kind described. We want to approximate, if at all possible, some solutions to the problem of making sense out of the interaction of two persons who are engaged in a face-to-face interaction. We can draw on existing knowledge concerning the analysis of language and at the same time try to include procedures that will help us to solve the problems that are of special relevance to face-to-face interaction. We recognize that we must include instructions to the computer so that the program can make the same interpretations of the language used by the participants as they themselves make. At one level, this involves discovering and programming that set of rules for analyzing language which is common to users of the same language. At another level, it involves analyzing the interaction in terms of the distinctive set of meanings which the particular individuals may be engaged in creating and developing through their continuing conversation. The rules that are built into a dictionary for assigning meanings to words, phrases or sentences represent a theory of the way in which these words and phrases are understandable to those who use them. In explicating these rules, we can draw on all that is known about the language in question (English). We can draw from the extensive knowledge that exists concerning grammatical and syntactic structure, for example, in developing sets of rules for interpreting whether a word is a subject, noun, or verb, whether a person is being talked about or talked to, whether an action is being described as having occurred or as about to occur, etc. Explicating general meanings, those known in common by "most" users of the language, is perhaps the easier (though none the less complex) task than explicating the idiosyncratic meanings which may be held by these and only these two persons.

The distinction here is one between the subjects' own world of meanings which serve as a framework for their making sense out of their continuing conversation, and the general set of meanings which any one of us as users of the same language would share. If we want to enter into their world of meanings so that the same sense which they make of their conversation can be understood by us, then we have to make some effort to understand their meaning structures, and their rules for "making sense". At the present time, this task seems to be more difficult than we can handle. We will be satisfied if we can capture the "common cultural meanings" and, by specifying these as rules, enable the computer to make similar interpretations. Those meanings which any one of us would make if we were in the same situation and in the roles being performed are the meanings that we will focus on.

We want to add the considerable information that comes from our knowing that the two persons are a patient and a therapist. We do not view them as just two people who are sitting and talking. Rather, we recognize that we are dealing with special roles and these roles carry with them some set of meanings that are properly designated as constituent elements of the role. How would anyone know that the person talking is a therapist? If he is playing that role, then there must be some patterns in his talk which are distinct and different from those of the patient or from those of any other kind of role found in a two-person interaction. He may show some patterns that are the same as those used by any speaker of the language since he is speaking the same language as others. However, we want to look for those patterns which are special or what might be termed the constituent elements of the role of, say, therapist. If we can find these, then we will be able to classify some of his talk in terms of the characteristic patterns that represent the therapist role. Similarly, with reference to the patient, we want to be able to classify some of his statements in terms of those key patterns that represent the patient role. In order to do this, we want to learn as much as we can about these two roles so that we can improve our chances of capturing the meanings of the statements generated by both parties in the interaction. Aside from an examination of the literature to learn what has been written about these roles, we want to examine actual verbal interactions, transcripts of therapy interviews, to determine what patterns of interaction via language may occur. Our task of developing a dictionary for these two roles can thereby be facilitated by examination of a set of data. That is, the task begins with the examination of interaction by persons in the same situation that we are studying.

MULTIPLE DICTIONARIES AND THE GENERAL INQUIRER

Given the characteristics of two-person interaction referred to above and given further that each participant has a different role to perform, multiple dictionaries are required to classify the content of the interaction.

Multiple rather than single dictionaries are needed in

order to make classifications of the content and tactics of interaction using different levels of interpretation, different units for classification and, in short, multiple modes of interpretation. General content dictionaries such as the Harvard III Psycho-Sociological Dictionary, could be applied to the same data for which special purpose dictionaries are also being used. Dictionaries designed to analyze selected aspects of the role of therapist and patient are one example of such special dictionaries. Any number of dictionaries, some overlapping with others, could be developed to tap particular dimensions of the content and tactics of interaction. How these different dimensions were interrelated would be a matter for further study but empirical determinations could be made if the several dictionaries could be applied to the same data and if all tagging was available for use in subsequent retrievals and analysis.

Heretofore, in the several uses of the GI system, only one dictionary had been applied to the set of data being processed. The dictionary could be fairly complex or relatively simple but the basic strategy of every dictionary was to assign tags either to single words, words and phrases (idioms), or sentences. Because of limitations of programs and hardware, it has not been possible to use multiple dictionaries, each of which might use a different unit as the unit for tagging, in processing data with the General Inquirer System. In addition, the GI system was oriented to written text.[3] As it was initially developed,[4] the GI allowed only word look-up so that tags were either assigned or not assigned to a word, depending on whether the word was in the dictionary or not. A Quick Dictionary in which were listed many frequently used words such as articles, determiners, prepositions, etc., were assigned no tag but by virtue of being included in the dictionary, the word would not appear on the leftover list. A major revision of the GI made it possible to tag idioms, that is, words which when used with other words would take on a different meaning from that of each of the words when used separately. Conditional tests, either for particular words or for tags within a specified range around a key word, became possible thereby extending the tagging system. However, it was not possible to include in the conditional test an examination of the ID field. Thus, the assignment of a particular tag could not be made conditional on the code in the ID field which might contain some reference to the characteristics of the source. Nor was it possible to test for final punctuation in a sentence and use the type of punctuation as part of the conditional. Eventually, Sentence Summary Tagging (SST) routines were added which made possible the assignment of tags to a sentence, depending upon occurrence of specified tags in a specified sequence.[5] Such tags could even be assigned depending on the pattern of tags occurring in sequential sentences. This expanded the flexibility of GI tagging but required a separate pass of the data after tagging had been completed. Further, in all programs written up until 1968, there was a limitation of 100 tags imposed on the size of the dictionary.

These limitations were of special significance for the analysis of dyadic interaction and only an expanded and revised GI system could solve them all. Goldhamer has incorporated many of these changes in a revised and expanded GI tagging system (GIT).[6] This version differs substantially from that described in the *User's Manual for the General Inquirer*,[7] but is still restricted to one hundred tags. Further, the IBM 7094 has a core load dictionary which makes the use of large (over 4,000 words) dictionaries difficult to process. The arrival of the IBM 360 System made possible new flexibilities and our experience with earlier versions of the GI suggested the necessity of others. Reference to some of these changes and the 360 system[8] will be made in this discussion to show how the analysis of dyadic interaction can be facilitated. In the main, we have been trying to solve the multiple dictionary problem with existing 7094 programs but our revision of the GI system, now called INQUIRER II, will solve the problem more directly and simply.

We have been proceeding with the development of several dictionaries in order to analyze the interaction occurring between therapist and patient. Three different dictionaries will be described and the kinds of results that may be expected from their application will be indicated.

INTERPERSONAL IDENTIFICATION DICTIONARY (IID)

An example of a person identification dictionary is the Interpersonal Identification Dictionary (IID) which we have developed for the analysis of the Wolberg case.[9] This

[3] *Ibid.*

[4] P. J. Stone, *et al.*, "The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information", *Behavioral Science*, 1962, 7, 484-498.

[5] These are reported by Ogilvie in his description of the Harvard Need-Achievement Dictionary. See P. J. Stone, *et al.*, *op. cit.*, 1966, pp. 191-206. Thus far, only tag tests are possible.

[6] Donald H. Goldhamer, "Toward a More General Inquirer: Convergence of Structure and Context on Meanings", paper presented to the National Conference on Content Analysis, Annenberg School of Communications, Philadelphia, 1967.

[7] Philip J. Stone *et al.*, *User's Manual for the General Inquirer* (Cambridge, M. I. T. Press, 1967).

[8] A more complete description and outline of the system is presented in D. J. Arp, G. Psathas, and J. P. Miller, "A Brief Introduction to the INQUIRER II: A Computer System for Contextual Analysis and Information Retrieval", mimeo, 1968.

[9] The Wolberg case consists of nine verbatim transcripts of a brief

type of dictionary, designed primarily to eliminate costly hand coding previously used to identify persons, classifies persons according to their relationship to the speaker. However, it requires information concerning the person named by the speaker since it does not know who "Mary" is when the name is mentioned. Personal pronouns such as *he* and *she* are more difficult to identify since their referents can change in the course of the interaction, though at least a classification of these as to sex role can be made. Self-references using personal pronouns such as *I*, *me*, *my*, are somewhat easier to classify since a check of who is speaking can indicate whether it is the doctor or the patient that is being referred to, though even this is not always accurate since indirect references can be made. Because such distinctions are important, two separate IID's are needed, one for each speaker. Thus, *John* may be tagged FAMILY-OF-PROCREATION + MALE when mentioned by the patient but tagged only MALE when mentioned by the doctor, since he is not a member of the doctor's family. The person taken as the point of reference in the IID determines how the classification is to be made.

In order to classify persons, it is necessary to have information concerning the identity of all persons named in the interview or series of interviews. Since all the data are available, a quick reading is sufficient to select all proper names and personal pronouns that occur in the data. Basically, the IID represents the construction of a table of equivalences such as: JOHN = HUSBAND, SAM = BROTHER, SUSAN = SISTER, etc. Each subsequent mention of the name is looked up in the dictionary and assigned to the appropriate category. It may be necessary to build context searches to make some classifications unambiguous, e.g., *my wife* is not to be classified the same as *his wife*. A backtest for particular pronouns may achieve this clarification. It is obvious that in the reading of the data to construct the IID, it is necessary to look for such context indicators so that conditional tests can be specified. Tags used in the patient IID for the Wolberg case are listed in Figure 1.

Tag No.			
87	LOWER STATUS	93	FAMILY OF ORIENTATION
88	EQUAL STATUS	94	FRIENDS
89	HIGHER STATUS	95	LOVE OBJECTS
90	MALE	96	PERSONAL AUTHORITY FIGURES
91	FEMALE	97	SELF
92	FAMILY OF PROCREATION	98	THERAPIST
		99	IMPERSONAL

Fig. 1. Patient Interpersonal Identification Dictionary

intensive psychotherapy of a female patient treated and reported by Lewis R. Wolberg in *The Technique of Psychotherapy* (New York, Grune and Stratton, 1964).

The immediate context does not always provide the indicators that are sufficient to classify persons. Consider the following example from a Wolberg interview to see how a named person can be identified in terms of his relationship to the speaker.

123 Pt. : And I was talking to *an old friend of mine*. *She*'s the one that recommended you. *She* said you helped *her* a lot and *she* was sure you could help me.
124 Dr.: You would really like to get rid of this trouble?
125 Pt. : Doctor, there is nothing I wouldn't do to get rid of it. Life doesn't mean anything, you know, the way things are going.
126 Dr.: How did you come to the conclusion that it was your nerves that were at fault?
127 Pt. : Well, doctor, you know *Mrs. Henshaw* and I'm very fond of her, and I've seen how she's come along so nicely that I thought that maybe I could get something out of it, too.

In 123, the patient indicates the nature of the relationship (an old friend) but does not name the person until 127 (Mrs. Henshaw). For the purposes of an IID, Mrs. Henshaw can be classified as FRIEND, FEMALE. (When the therapist refers to Mrs. Henshaw, the same tags could be applied so long as it is clear that the relationship is defined with reference to the patient.)

If an IID were not available, the program would have no way of knowing that this word is the name of a person rather than just another word. Capitalization is a common way of indicating proper names in writing, but if capitals are not available or not used in inputting data, then this device is not possible. Even if capitals were used, how is Henshaw to be distinguished from Florida — both are names, but one of a person and the other of a place — unless all possible names were to be included in the dictionary? One might note the presence of *Mrs.* in front of *Henshaw* as a clue to a person. Thus a backtest for titles of address could conceivably resolve the ambiguity of whether "Henshaw" is a word or a person. The problem would remain if a first name, *Mary*, were used to refer to her instead of a last name. And what if the last name were used without a title preceding it?

One approach to solving the problem is to provide the computer with a way of distinguishing names of persons from places and then asking "who is that?", or "what is that?", or "where is that?" In interacting with the content analyst, the computer could wait for this clarification before proceeding. Otherwise, it would put this word on the leftover list together with all others not found in the dictionaries.

An alternative would be to provide rules for searching the context, possibly extending back to the preceding state-

ments, to reduce the ambiguity concerning who Mrs. Henshaw is.

In the example given, in 123, the patient's second statement refers to an old friend as *she*. To classify the pronoun *she* in terms of more than the tag FEMALE, a backtest to the preceding sentence is necessary. When the word *friend* is found, the tags FEMALE *and* FRIEND could be applied. This procedure would not necessarily be correct if the sentences were "...my mother's old friend. She was a great help to me..." since it would not be clear from a backtest whether *she* referred to "mother" or to "mother's friend". The elaboration of the rules for making such disambiguations remains a major problem. We wish to point to the possibility of solutions for this problem if procedures similar to Stone's disambiguation rules were followed. The problem of defining relationships for personal pronouns may be approached in a manner similar to the clarification of word-senses. An analysis of the context in which persons' names occur should yield a set of rules for distinguishing persons from places.[10]

It can therefore be seen that the construction of an Interpersonal Identification Dictionary (IID) involves not only the development of a set of tags defining the types of relationships which are of particular interest to the investigator but also the development of procedures for disambiguating whether a word refers to a person and what that person's relationship to the participant in the conversation is. The basic notion behind an IID is similar to that found in preliminary stages of interpersonal relationships namely, identifying particular persons so that "we", the persons involved in the dialogue, know who we are talking about. Persons are then known to "us" when subsequently referred to. The specific sense of a conversation involving named others depends on our sharing the same set of meanings about who these other persons are. Separate IID's involve, in addition the notion that my view is different from your view and that others do not stand in the same relation to me as they do to you, though they may, for certain purposes, be the same to "us". Elaborating these special distinctions is important. We could also note that the same problem is involved in developing special meanings for words which we use in a developing conversation, i.e., the creation of a common culture. What words mean to "us" as we use them may be somewhat different from what they mean to others. We may want to increment a general dictionary with those specific meanings which our words have for us. If so, the problem of incrementing and modifying a dictionary of words is similar to that of defining names for an IID, as outlined above. Names of persons are most similar to those kinds of words which, although they exist in the language, can be defined differently for each person depending on his relationship to the person named. The general problem is similar to that of learning programs, i.e., programs that can acquire a new concept and add it to that which they already "know". Some information about the new name must be obtained before it can be assimilated into the set of meanings (dictionary) already in use or have a new meaning (tag) added to the dictionary.

[10] The procedure would be similar to the strategy reported by Stone in disambiguating different word-senses. Philip J. Stone, "Improved Quality of Content Analysis Categories: Computerized Disambiguation Rules for High Frequency Words in the English Language", paper presented to the National Conference on Content Analysis, Annenberg School of Communications, Philadelphia, Pennsylvania, 1967.

THERAPIST TACTICS DICTIONARY (TTD)

Another type of dictionary is that which tries to classify the tactics of interaction rather than the content. A different unit, such as the sentence, utterance or that burst of speech sandwiched between the other person's preceding and succeeding remarks may be the unit to be classified. For example, one may wish to classify statements according to whether they are interrogative, exclamatory, declarative, etc. Such classifications may then be tabulated in order to describe some aspects of a person's interaction style. In the need achievement analysis, Ogilvie examined sequences of sentences in order to determine the presence of achievement imagery.[11] Similarly, a therapist's statement can be examined in order to classify it in terms of interaction strategies or tactics such as the therapist's use of open-ended probes.

An example of such a dictionary is illustrated by our development of a Therapist Tactics Dictionary (TTD) which is designed to examine and classify the tactics used by the therapist. Content patterns are determined by the presence of combinations of words and/or tags, the sequence in which they occur in the sentence and the presence or absence of other formal markers which designate whether the sentence is a question or statement, positive or negative.

It is also possible to define some tactics on the word level. For these, a tag is constructed which contains all words identified as belonging to that tactic. An example is the tag DIRECT-PRAISE which includes words and phrases such as *fine*, *good*, *very well*, *excellent*.

This dictionary developed out of our previous work[12]

[11] Ogilvie, *op. cit.*

[12] G. Psathas, and D. J. Arp, "A Thematic Analysis of Interviewer's Statements in Therapy Analogue Interviews", in P. J. Stone, *op. cit.*, 1966, pp. 504-523; and D. J. Arp, and G. Psathas, "Verification and

analyzing therapy-analog interviews using the earlier, 1962 version, of the GI. Retrievals were made for subject of the sentence and the presence of particular words. These retrievals were used to build new tags which could then be used in subsequent retrievals.

The strategy in the Therapist Tactics Dictionary (TTD) is more flexible since the 1967 GI is used including Sentence Summary Tagging (SST). A dictionary has been devised (see Figure 2) which includes the feature of tagging some words according to their grammatical function and others according to their relevance for therapist tactics. The first part of the dictionary serves a function similar to the marker words which Stone has described in his disambiguation project. In fact, many of our tags are similar since we had exchanged ideas during the early stages of this work although we had arrived somewhat independently at a list of necessary markers.

01 Determiners (2-15)
- 02 Articles
- 03 Demonstratives
- 04 Demon. I
- 05 Demon. II
- 06 Possessives (7-10)
  - 07 2nd Person
  - 08 1st Single
  - 09 1st Plural
  - 10 3rd Person
- 11 Numbers
  - 12 Numbers Cardinal
  - 13 Numbers Ordinal
- 14 Negative (*no, none*)
- 15 Prearticle

16 Pronouns (17-31)
- 17 Personal (18-28)
  - 18 Non-Reflexive (19-23)
    - 19 *You*
    - 20 *I, me*
    - 21 *We, us*
    - 22 *Others* I
    - 23 *Others* II
  - 24 Reflexive (25-28)
    - 25 *Yourself*
    - 26 Reflexive *I*
    - 27 Reflexive *We*
    - 28 Reflexive *Other*
- 29 Impersonal (30-31)
  - 30 Non-Reflex. Impersonal
  - 31 Reflex. Impersonal

32 Auxiliary Verbs (32-39)
- 33 *To be*
- 34 *To do*
- 35 *To have*
- 36 *To be able to*
- 37 *Should* (*shall, should, ought*)
- 38 Volition (*will, would*)
- 39 *Must*

40 Adverbs (41-42)
- 41 Time
- 42 Positional-Spatial

43 Direct Praise
44 Openers
45 Tentativity
46 Question Triggers
47 Special Verbs (48-51)
- 48 *Tell*
- 49 *Remember*
- 50 *Think*
- 51 *Feel*

52 Prepositions
53 Conjunctions
54 Not (*not*)
55 Space Reference
56 Time Reference
57 Quantity Reference
58 Quality Reference
59 Emotional State
60 *Continue*
61 *Begin*
62 *Want*
63 *Mean*
64 *Let*
65 Mild Agreement
66 Regular Agreement
67 *Know*
68 *Summarize*
69 *Compare*
80 Ability Potential
81 Open Ended Modifier

Summary Tags (84-93)
- 84 Direct Question
- 85 Statement
- 86 Negative
- 87 Positive
- 88 Probing Reflection
- 89 Prodding Suggestion
- 90 Direct Urge
- 91 Direct Praise
- 92 Agreement
- 93 Disagreement

Fig. 2. Therapist Tactics Dictionary (TTD)

As Figure 2 shows, Auxiliary Verbs are divided into seven sub-classes (*be, do, have, able, should*, volition and *must*). Each sub-class may contain one or more verbs. It is thus possible to instruct the computer to distinguish between the statements *you should do X*, and *you have done X* by referring to the appropriate sub-class of auxiliary verbs.

To illustrate how the TTD may be used to classify a sentence, consider the followings types of statements made by the therapist which we call PROBING.

PROBING
What did you think of that?
You mean you think that is necessary.
Then you are saying in all these ways you are different from him.
Or are you telling me that is what he thought?

A number of these sentences were collected from our earlier study and examined to determine what patterns of words and/or tags, in what sequence and within what range of one another could be found. Four basic types of PROBING statements were defined. These were defined as follows:

Probing Type

I IF (OCCUR1(33,0).AND.OCCUR3(81,0,2))

II IF ((OCCUR3(19,0,4).OR.(OCCUR1 (44,0).AND. OCCUR2(19,0))).AND.OCCUR2(47,0))

III IF ((OCCUR3(19,0,4).OR.(OCCUR1(44,0).AND. OCCUR2(19,0))).AND.OCCUR2(32,0).AND.OCCUR2(47,0))

IV IF (OCCURI(30,0).AND.OCCUR2(63,0).AND. OCCUR2(19,0))

The sentence summary tagging is written in FORTRAN IV and the instructions are defined in the *User's Manual for the General Inquirer*.

Basically,

OCCUR1 (TAG, SYNTAX-CODE) starts the scan at the beginning of the sentence and tests for the tag indicated.

OCCUR2 (TAG, SYNTAX-CODE) continues the scan and tests for the tag indicated.

OCCUR3 (TAG, SYNTAX-CODE, RANGE) continues the scan for words within the range.

Other instructions of the SST are also available but are not listed here.

Let us take Type I PROBING to see how it works. Type I: if tag 33 (an auxiliary verb, *to be*) regardless of its syntax code, is found and is followed by tag 81 (an open-ended modifier) one or two words later, the sentence is classified as PROBING. If this test fails, then the next one is attempted.

Replication of a Thematic Analysis of Interviewer's Statements in Therapy Analogue Interviews", paper presented at the meetings of the Midwest Psychological Association, 1967. The dictionary described here was first programmed to run with the sentence summary feature of the 7094 system by Robert Miller.

Is there anything else that you can remember?
33 81

Are there some other things you can recall?
33 81

Type III PROBING would be found if [tag 19 (YOU) were found within the first four words of the sentence] or [if the sentence began with tag 44 (OPENER) and were followed by tag 19 (YOU)] and then tag 32 (any AUXILIARY-VERB) and tag 47 (any SPECIAL-VERB)] were found.

Examples:

Then you are saying in all these ways you are different from him.
44 19 32 47

You were telling about his episode.
19 32 47

The tag order as well as tag occurrence is considered. In examining the way the language is used, it often turns out that sequences determine certain word usages. For example, the syntactic structure of the sentence is in large part determined by the sequence. Thus, if "YOU" occurs in the above two sentences, in either the first or second word position, it is probable that it will be the subject of the sentence. By this kind of detailed analysis of sentence structure and the construction of a set of specifications, we begin to detect patterns of usage of the language that we had not noticed before and some of our classification problems are thereby solved.

As the interaction proceeds, it is desirable to not only note the number of Probes generated by the therapist but also to note the types of patient statements that follow these. In order to do this, it is necessary to make some classification of patient statements either in terms of general content or of patient tactics. We are engaged in formulating a Patient Tactics Dictionary (PTD) which would classify the patient's statements into categories including many which have been of major theoretical concern in the study of psychotherapy, e.g., resistance, ambivalence, intellectualizing, etc. Although these categories are generally defined more broadly and take a larger context into account, we are looking for operational equivalents on the sentence level.

The TTD and the PTD are separate dictionaries which are looked up depending on which person is speaking. Prior tagging by a general purpose dictionary is assumed since some conditional tests using these dictionaries are based on tags which have previously been assigned. Thus there is a sequence of tagging involving several dictionaries. One of the first would be a dictionary that assigns tags according to their grammatical function (see the first forty-two tags in the TTD, Figure 2). Another would be a general psychological content dictionary. Both of these are general in scope with tags assigned by either or both dictionaries available to be used in conditionals in succeeding dictionaries. The strategy for tagging would be to run those dictionaries that are dependent on earlier assigned tags last. Such a multiple-pass processing of multiple dictionaries is another method for achieving what has long been sought in GI tagging, namely forward tests. Heretofore, tagging conditional on tags or words that *followed* the key word was not possible. (Goldhamer has recently added this feature to the GI 7094 version by use of elaborate iterative procedures.)

THE PSYCHOLOGICAL CONTENT DICTIONARY (PSYCODIC)

The Psychological Content Dictionary (PSYCODIC) is yet another dictionary involved in the analysis. The list of tags which it contains is presented in Figure 3. Because this dictionary is similar in basic philosophy to the Harvard III Dictionary for Psycho-Sociological Content, less space will be devoted to its discussion.

This dictionary is designed to classify words and phrases occurring in the therapy interview situation according to general social psychological theory. The level of meaning is the more overt or denotative level but not necessarily that which the patient himself would adopt. It is closer to the meanings which the therapist would use as he interprets the psychological significance of the content. It was developed after the examination of Key-Word-In-Context print-outs of many therapy interviews, and therefore is able to classify the content of verbal interaction better than those dictionaries developed for the analysis of written text. It is designed to classify both the patient's

Tag 01 Age and/or Time
Tag 03 Body Parts
- Tag 04 Internal Body Parts
- Tag 05 External Body Parts
- Tag 06 Genital Body Parts

Tag 07 Somatic Conditions
- Tag 08 Health
- Tag 09 Illness
- Tag 10 Death

Tag 11 Treatments
Tag 12 Tests
Tag 13 Non-specific References
Tag 14 Object Orientation
- Tag 15 Retaining
- Tag 16 Expelling
- Tag 17 Attaining

Tag 18 Gender
- Tag 19 Male
- Tag 20 Female
- Tag 21 Neuter

Tag 22 Authority Figures
- Tag 23 Treatment-Guidance
- Tag 24 Legal

Tag 25 Sensory Perceptions
- Tag 26 Visual
- Tag 27 Auditory
- Tag 28 Gustatory
- Tag 29 Olfactory
- Tag 30 Dermal

Tag 31 Action Norm
- Tag 32 Passive
- Tag 33 Active

Tag 34 Overt Sexual Acts
- Tag 35 Sexual Action
- Tag 36 Fore-Play
- Tag 37 Pre-Fore-Play

Tag 38 Social Action
- Tag 39 Aggressive
- Tag 40 Friendly
- Tag 41 Isolative
- Tag 42 Communicative

Tag 43 Dominative
Tag 44 Submissive
Tag 45 Helping
Tag 46 Action Direction
Tag 47 Approach
Tag 48 Avoid
Tag 49 Action Achievement
Tag 50 Success
Tag 51 Failure
Tag 52 Cognitive Processes
Tag 53 Contemplative
Tag 54 Cognitive Awareness
Tag 55 Uncertainty
Tag 56 Decisiveness
Tag 57 Regard
Tag 58 Positive
Tag 59 Negative
Tag 60 Fortune
Tag 61 Good
Tag 62 Misfortune
Tag 63 Emotional Type
Tag 64 Anger
Tag 65 Affection Present
Tag 66 Affection Absent
Tag 67 Fear and Apprehension
Tag 68 Happiness
Tag 69 Sadness
Tag 70 Distress and Arousal
Tag 71 Positive Emotional Value
Tag 72 Negative Emotional Value
Tag 73 Undefined Need States
Tag 74 Obstacles
Tag 75 Present
Tag 76 Struggle Against
Tag 83 Self
Tag 83 Specific Others
Tag 99 Non-Specific Others

Fig. 3. Tag Categories of the Psychological Content Dictionary

and the therapist's verbal statements and to facilitate the development and subsequent testing of hypotheses concerning the content of interaction in psychotherapy. In subsequent studies, it would be possible to compare the results from the use of this standardized psychological content dictionary (PSYCODIC) with either general or special purpose dictionaries applied to the same data.

The PSYCODIC incorporates the multiple dictionary principle by use of sub-sets of tags. A sub-set of tags consists of a header tag, called a super-tag, and one or more sub-headings, called mini-tags. An entry word (or phrase) may be assigned to one or more super-tags and to one or more mini-tags on a list. For example, consider the two tag lists 38 SOCIAL ACTION and 34 OVERT SEXUAL ACTS.

Super-tag	34 OVERT SEX	38 SOCIAL ACTION
Mini-tags	35 SEXUAL ACTION	39 AGGRESSIVE
	36 FORE-PLAY	40 FRIENDLY
	37 PRE-FORE-PLAY	41 ISOLATIVE
		42 COMMUNICATIVE
		43 DOMINATIVE
		44 SUBMISSIVE
		45 HELPING

The following entry words are assigned as follows:

hug = 34, 37, 38, 40
molest = 34, 36, 38, 39
date = 34, 37, 38, 40
embrace = 34, 37, 38, 40

Each of these entry words is assigned the super-tag (underlined) associated with the list. Each is also assigned one or more mini-tags as necessary. The general principle then is that every word is assigned a super-tag and only in instances where a word may not be clearly defined by one or more mini-tags will it be assigned no mini-tag at all. Entry words on one list do not necessarily have to appear on any other tag list.

In certain respects, each super-tag list can be regarded as a separate dictionary. The fact that all are included in what is called one dictionary is an indication that they are designed to be used together and that some interrelationships exist among the tags. As an example, in the case of the PSYCODIC, we are interested in determining relationships between tags such as 38 SOCIAL ACTION, 52 COGNITIVE PROCESSES and 63 EMOTIONAL TYPE. The frequency and patterning of co-occurrences can be examined for the set of data being analyzed. Since the same dictionary is being applied to both participants in the interaction, the therapist and the patient, it is possible to determine similarities and differences between them. We can try to determine the extent to which they are talking about the same things. That is, if patient statements include co-occurrences of ANGER and UNCERTAINTY, do the therapist's statements also show the same patterning? Do similar proportions of the sentences generated by each contain the same tags? Are co-occurrence patterns similar within the sentences generated by each participant? For example does the therapist's pattern more frequently contain ANGER + COGNITIVE AWARENESS? Does this indicate that he is trying to get the patient to think about and become more aware of his feelings of hostility? Validating interpretations of patterns will depend on retrievals done specifically for that purpose. The PSYCODIC offers the possibility of interpretations relating to the psychological significance of the interaction and to the assessment of changes in the patient's functioning as these are reflected in the content of his talk.

PROCEDURES

I now want to examine some of the system requirements and specific procedures involved in the kind of multiple dictionary look-up being described.

1) The source of a statement must be identified in order that the appropriate dictionary(ies) will be searched. For example, if the speaker is the therapist then the Therapist Tactics Dictionary rather than the Patient's will be searched. It is also possible for the assignment of tags within any dictionary to be conditional on the particular characteristics of the speaker that may be coded into the ID field, such as age, sex, social class, demographic and psychological characteristics. For example, if the speaker is a young child, references to Family will more likely be Family of Orientation.

2) Some tags may be assigned conditional on words

and/or tags in the previous sentence or on the utterance generated by the previous speaker. For example, the distinction between pro-action and reaction made by Bales[13] in scoring interaction was in terms of whether the previous statement was spoken by a different speaker (reaction) or by the same speaker (pro-action). In analyzing the relation between therapist tactics and succeeding patient statements in order to assess the effects of particular therapist tactics, a check for the source of the previous statement is necessary.

Depending on the content and the source of the previous statement, particular tags may be assigned. For example, if the previous utterance was by a different speaker, and if it included particular content references, then Tactic A may be assigned. However, if the previous utterance was by the same speaker on the same content topic, then Tactic B may be assigned. Thus, it becomes necessary to temporarily store the previous statement in order to allow a comparison to be made at the time the next statement is ready for classification. In this way the source and the content of the statement can be examined to determine whether conditional specifications have been met.

However, we are aware that, in interaction, the meaning of a statement may depend on a whole series of prior statements as well as on subsequent statements. Thus it would be ideal if several statements and not just the preceding could be examined before the next one was tagged. Forward checks pose more difficult problems since the statement must be stored until subsequent statements are examined. For example, a *how are you* spoken by the therapist may be taken as a greeting or as a question about one's state of health. How it is taken by the responding other cannot be known until he responds. If his response is *fine, how are you*, then the original statement may be classified as a greeting. If he answers at length about his health, it may be taken as a probe concerning health. Among the more common solutions in computer content analysis has been to classify it as having both meanings or if that is not satisfactory to select the more frequent meaning. In interaction, however, the crucial question is what sense is made of the statement by those who hear it and respond to it. Its meaning lies in its use, i.e., the work that it does in the context of the interaction. This can be determined by observing its antecedents and its consequents. The response of others tells us how they are interpreting it and what it means to them in the context.

The prospective-retrospective nature of meaning cannot be handled to our satisfaction unless forward *and* backward conditionals are possible. If only retrospective conditionals are possible, then the intent of the speaker must be guessed and used as the major basis for classification. If prospective conditionals were possible, then the response of the other could be used together with the guessed intent of the speaker in deciding on the classification. A possible solution to this is what might be called "tentative tagging". That is, tags are assigned but held subject to change after some specified sequence of next statements. In the above example, if the discussion of health and physical symptoms followed the *how are you*, then the greeting could be re-tagged as a health probe. Obviously, the specification of which tags are tentative and for how long they are to be so regarded involves the investigator in the elaboration of his theory of meanings. In fact, the entire set of procedures being outlined here represent a direct confrontation of the issues involved in determining how humans actually do interact and how they themselves assign meanings to communications.

3) Because the number of tags assigned in such multiple dictionaries is enormous, some means for reducing these numbers in making summary tabulations is necessary. For example, super-tags, which represent major content categories, can be used for summary purposes rather than mini-tags, which are sub-categories. We have used these distinctions in the TTD, and the PSYCO Dictionary, for example, in order to be able to examine major category frequencies before deciding whether to look within the category to its various sub-classes. For example, the super-tag PRONOUNS can be examined to determine what its frequency is before looking within it to sub-distinctions between types of pronouns, e.g., PERSONAL, IMPERSONAL. In the PSYCO Dictionary, the super-tag SOCIAL-ACTION can be tabulated without showing a breakdown for AGGRESSIVE, FRIENDLY, COMMUNICATIVE, etc. For both super- and mini-tags, it is possible to make comparisons with specified comparison groups such as base rates, previous sessions, or ideal norms. For example, it should be possible to determine what topics are being "avoided" as well as which ones are "high" in comparison with specified norms.

Contingencies and inter-relationships between content categories can be calculated and reported. Thus, the association of particular content themes with each other within a specifiable unit such as an interview or over several sessions could be calculated. For example, has there been any change in the reporting of symptoms, of relationships with particular other persons, or in the descriptions of past experiences. Indices of change can be reflected in the interrelationships between tags over time.

[13] R. F. Bales, "The Equilibrium Problem in Small Groups", in T. Parsons, R. F. Bales and E. A. Shils, *Working Papers in the Theory of Action* (Glencoe, Illinois, The Free Press, 1953).

4) A comparison of the interaction of two speakers is necessary such that a matrix can report the kinds of statements (tactics or content) generated by one speaker which are followed or preceded by particular statements by the other. In keeping track of what is classified, it should be possible to indicate, for example, whether patient intellectualization follows particular therapist tactics, such as interpretation or open-ended probes more frequently. The requirements for making such a classification are different from those of an ordinary tag tally or sentence retrieval system. In this instance, a matrix operation of some kind is needed such that for every therapist tactic a count is kept of what patient tactic follows. We can call this an "interaction map". A similar notion was involved in Bales[14] tables of pro-action and reaction in which he tried to show which Interaction Process Analysis category followed another category. In this way, the probability that one type of category would follow another could be empirically described.

The same kind of interaction mapping should be possible for the general content of the interaction as classified by the PSYCO Dictionary. The question here is whether certain content patterns go together. For example, are they "talking about the same thing" or are they diverging,

CONCLUSION

Considering the features of two-person interaction presented initially, the strategy of multiple dictionaries for special purposes and including different types of units for tagging represents an approach to the special problems posed by this kind of data. We do not feel that all issues have been satisfactorily solved. With the advent of new hardware in the form of the IBM System 360 and the development of more flexible software which will allow such procedures as forward tests as well as back tests, larger dictionaries with a larger number of tags, as well as a larger number of dictionaries, multiple passes of the same data so that dictionaries used in a second pass can include conditional tests for tags applied in a previous pass, etc., some solutions have been achieved.[15] Contributions to the extension and development of the General Inquirer have already resulted in the major re-programming of the system and a new designation as INQUIRER II.

Among subsequent extensions and revisions that can be anticipated is the analysis of interaction as it is generated in an interactive mode, on-line to the computer. The interaction may be between two persons at remote in-put devices or between a person and a stored program in the computer. The aim of such a system would be to analyze interaction in-process, i.e., as it occurred. Classifications of the content would be made according to various stored dictionaries, summary tables could be made, statistics computed, and feedback provided to one or both of the participants. The system would take on some of the characteristics found in teaching programs, e.g., computer aided instruction. In order to allow for increments to stored dictionaries, particularly dictionaries such as the IID described here, some features of learning programs need to be included. That is, the system would have to be able to interrogate the user in order to obtain information concerning how new words, not previously placed in a dictionary, could be tagged. The classification and analysis of interaction as it proceeds presents a more challenging problem but one which can be solved as computers expand in capacity and as time-sharing becomes more and more feasible. Solutions to the problems of analyzing interaction, as described in this paper, can contribute to the further extension of the system so that data generated in on-going interaction can also be analyzed.

[14] *Ibid.*

[15] Each of these requirements is being incorporated into the 360 system (INQUIRER II) in order to facilitate the analysis of interaction. They will extend the capability of the system for handling conventional text materials as well. The 360 system will run under IBM's Operating System (OS) with PL/I. Dictionaries are contained on disk, drum or core and a maximum of 256 × 256 = 65,536 tags (enough to accommodate any present user) can be used. Data tagged in one pass can be used as input in another tagging run. The context that can be checked in deciding on what tags to apply can include any or all of the following: backward and forward checks for words and tags; sequence of words and tags; terminal punctuation; ordinal position of the word within a sentence; characters in the ID field; and the syntax assigned to a word or tag. See the detailed description of the system in D. J. Arp, G. Psathas, and J. P. Miller, *op. cit.*

Some Long and Short Term Trends in One American Political Value: A Computer Analysis of Concern with Wealth in Sixty-two Party Platforms*

J. ZVI NAMENWIRTH
Yale University

The question whether value change causes social change or vice-versa can only be answered by historical inquiry. In an attempt to answer this question this study applies quantitative procedures. Content analysis of Republican and Democratic party platforms over a hundred-and-twenty year period produced campaign-to-campaign fluctuations in concern with wealth values. These raw scores were decomposed into long-term trends and short-term deviations from these trends which were subsequently related to economic indicators. In the long run, both parties move closely together, although the Democrats increase their concern at a greater rate than the Republicans. Furthermore, both parties respond in similar manner to changes in the economic environment. Regarding short-term changes, the relationship between the parties is more complex. While in the early period (1844-1896) the parties move in opposite directions from campaign to campaign, in the later period (1896-1964) they tend to move in similar directions. Also, while the Democratic party responds immediately to short-term changes in the economy, the Republican party does not or only weakly so. To conclude, economic changes precede rather than follow values changes in these platforms.

INTRODUCTION

Content analysis of American party platforms is a convenient way to assess changing magnitudes in a variety of political and social values. The information produced shares a "small talk" quality with much content analysis research and the bulk of historical inquiry: people like to know what others do, think, feel and want. The distance of time, furthermore, imbues historical findings with an aura of mystery which seems, in turn, embedded in a yearning for descriptive knowledge of the past.

I, for one, have little patience with descriptive studies, particularly those in the area of values. My interest is abstract and theoretical. It is centered on a perennial issue in philosophy and social theory: do changes in values cause changes in the social environment, or do they merely reflect social changes brought about by the internal dynamics of the social process? Where so many thoughtful men have held conflicting points of view, there is little reason to expect that the truth is either simple or unambiguous. Indeed, there is little reason to expect that the causal relationships between values and social change are the same for all times, all values or all social processes.

Pondering these issues, Harold Lasswell and I (forthcoming) came to believe that for social change to cause value change, the act must precede the thought, while for value change to affect the social order, the reverse must be true. Let me elaborate.

Aggregate social processes often cause social change. For instance, population growth, a favorite example in this line of argument, will change the per capita distribution of social goods and it will often do so differentially for the different age groupings, the sexes, the various income strata and other pertinent political groupings. The principles and mechanisms of the precipitant causes are rarely understood or even perceived by contemporaries. Nonetheless, the resulting social changes may produce severe dislocations in the distribution of social goods in terms of existing expectancies. These dislocations then cause changes in the various conceptions of what the ideal production, allocation and distribution should be.

On the other hand, for value change to cause social change it must redirect the allocation of the social goods and therefore steer the political mechanisms which control these allocations. But in order to redirect the production

* Paper read at the National Conference on Content Analysis, Annenberg School of Communications, University of Pennsylvania, November, 1967. — I am indebted to the National Science Foundation for support of this research (GS-00614-B), to M. Harvey Brenner for advice on time trend analysis, to Harold D. Lasswell for the constituent value formulations, to my research staff and especially to John R. Hall for many innovative ideas and procedures. This study will be published in a forthcoming book of congress papers, John Wiley and Sons.

and distribution of goods, the decision process must gain control of the material as well as the social environment. Thus it follows that the increasing knowledge of the dynamics of things material and social and the subsequent mastery over these environments would make social change increasingly sensitive to changes in the value structure of society. At the same time, however, an increasing ability to gage social preferences would tend to restrict leaders in their freedom to manipulate the social system. Therefore, the net effect of these interactions between value change and social processes is not immediately clear.

These conceptualizations stress the mediating role of the political process in the nexus between value and social change. They explain the selection of party platforms for the study of value change and its social causes and consequences (Namenwirth and Lasswell, forthcoming), while predicating many other features of the analysis. Before advancing with this analysis, let me first define some of the basic concepts and describe the ways in which they were measured.

The distinction between goods and values is basic to the whole enterprise. Goods are the available resources of a society at any one time and these resources are not restricted to material and economic resources per se, but include the whole range of scarce desirables such as friendship, recognition, health or power. The aggregate sum of resources, their production, allocation and distribution and especially the inequality of this distribution across pertinent social groupings constitutes the social structure of a society at one point in time. Social change, therefore, connotes a change in this structure of goods and its attributes over a period of time.

Values, on the other hand, are goal states, conceptions about the desirable (rather than actual) level of goods and their production, allocation etc. in some future society. Value change implies changes in these conceptions of goal states and their hierarchical organization. Lasswell distinguishes between eight classes of values: power, rectitude, respect, affection, wealth, well-being, enlightenment and skill (Lasswell and Kaplan, 1950). My associates and I compiled over the past few years the Lasswell Value Dictionary, which contains at present seventy-six categories, most of them sub-categories of these aforementioned eight classes of values. From these, I selected for this paper, the *Wealth* value which was defined as "services of goods and persons accruing to the individual in any way whatever".

The *Wealth-other* category, my present value index, is a residual category, excluding the subcategories *Wealth-participants* and *Wealth-transaction*, and including a great variety of words such as *affluence*, *capital*, *currency*, *factory*, *livestock*, *steel* and *unemployment* — 195 words in all. Consequently, the residual category is poorly defined, but this lack of precision (if not validity) of the category is offset by a gain in reliability, a not uncommon dilemma in content analysis. To measure value change, I counted words, but as this is a favorite pastime in content analysis, I need not dwell on it here.

To measure social change in material goods, I used social indicators collected by various agencies of the U.S. government and recorded in *Historical Statistics of the United States* (U.S. Bureau of the Census, 1960; 1965; data for more recent years were obtained from the *World Almanac* and the National Industrial Conference Board's *Economic Almanac, 1964*).

Finally, a common and untested assumption of the trade predicates the entire study. The assumption maintains that differential occurrence of the category *Wealth-other* from document to document presents a precise measure of differential concern with the category, while this concern in turn is an appropriate measure of the relative priority of the value in the total value scheme of each and all documents.

LONG-TERM TRENDS

Using standard General Inquirer procedures (Stone, *et al.*, 1966), I computed the percentage frequencies of the category *Wealth-other* for the platforms of the two major parties for each presidential election from 1844 to 1964, sixty-two platforms in all (Porter and Johnson, 1961; *One Nation, One People*, Democratic Platform Committee, 1964; *New York Times*, July 12, 1964, p. 56; July 13, 1964, p. 20). Figure 1 presents these frequencies.

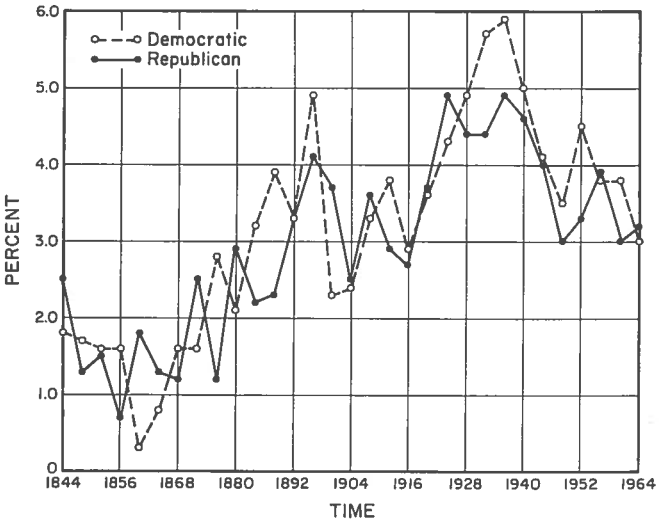

Fig. 1. Percentage frequencies of *Wealth-other* concern in Democratic and Republican party platforms, 1844 to 1964.

A number of regularities are immediately evident. In the first place, later platforms tend to devote more attention to the category than earlier platforms. Secondly, this linear tendency is interrupted by a number of peaks and valleys, the major positive deviations occurring in 1896 and in 1936. Finally, both parties' platforms tend to follow these patterns in a similar fashion, although the average Democratic concern exceeds that of the Republicans. The magnitude of such general trends is so overbearing that many smaller and less pronounced regularities become overshadowed. For this reason, it is advisable to decompose the *Wealth-other* frequencies into long-term trends and campaign-to-campaign deviations from those trends.

Two linear regressions of the *Wealth-other* categories over time represent the long-term trends while making the fewest possible assumptions about those trends. Figure 2 supports the earlier observation that, on the average, concern with *Wealth-other* increases over time.

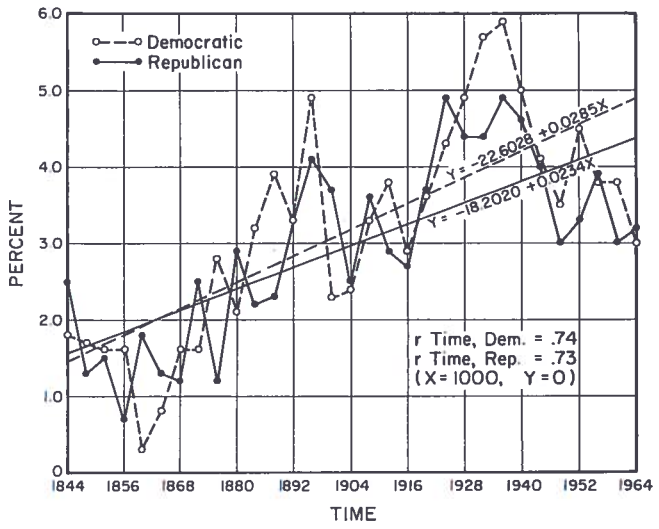

Fig. 2. Linear regressions of time and *Wealth-other* concern in Democratic and Republican party platforms, 1844 to 1964.

Furthermore, the rate of increase (slope) is greater for the Democratic platforms than for those of the Republicans.[1] Thus, Democratic concern begins to exceed Republican concern about 1860, and it does so at an increasing rate. For the entire period, 1844 to 1964, the Republican party platforms lagged an average of about eight years behind the Democratic platforms.[2] The above differences notwithstanding, the two trends are, of course, highly correlated.

The estimation of long-term trends of these data by simple linear regressions has a number of theoretical and empirical limitations. At some point in time, the increase in *Wealth-other* concern must reach an upper limit determined by the semantic and grammatical structure of language. It would be impossible to write a political document containing only words classified as *Wealth-other*. For similar reasons, there must also be a lower limit to these linear trends. Therefore, linear change is possible only within a limited time range. Inspection of the actual observations indicates gross and systematic deviations from the linear trends. For all these reasons, the long-term trends were also determined using a procedure of moving averages.[3] Figure 3 presents this approach to *Wealth-other* non-linear trends.

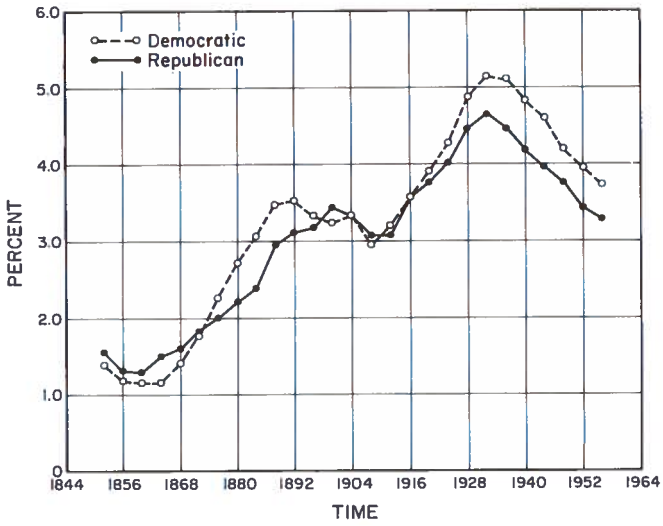

Fig. 3. Twenty-year moving averages of *Wealth-other* concern in Democratic and Republican party platforms, 1852 to 1956.

While not contradicting any of the observations made concerning the linear trends, the trends described by moving averages present additional information.

The upward trends in concern with *Wealth-other* is not continuous, but shows two peaks followed by a decline. The peaks occur in the last decade of the nineteenth century and around 1932 — both periods of severe economic dislocation. Furthermore, since the height of the depression of the 1930's, the average concern with the

[1] Similar findings hold for other values with linear trends. When the trend is positive (e.g., in the cases of concern with *Well-being-somatic* and the residual *Skill* category), the Democratic slope is greater than the Republican; when the trend is negative (e.g., in the *Rectitude-scope* and *Respect-indulgence* categories), the Republican concern decreases at a greater rate than does that of the Democratic platforms.

[2] The lag was calculated by taking the difference between the year the Democratic linear regression reached its average level and the year the Republican linear regression reached that same level.

[3] The five entry moving average used here takes the mean of *Wealth-other* concern in the first through fifth election years, the second through sixth years, and so on.

category has steadily declined (Stone *et al.*, 1966: 373). Finally, the most striking feature of these long-term trends of the two parties is their similarity. Indeed, the correlation between these two trends is .98, approaching identity. It is, however, one thing to describe this relationship and quite another to account for its magnitude.

Let me briefly distinguish between three rather different theories. The idealist theory would argue that immanent cultural processes produce value changes which will be reflected in the components of society, in this case, the political parties and their platforms. The similarity in the changes in value concern in the two parties is therefore a reflection of the changes in societal values, which result from some immanent process of spiritual growth and decay. A utilitarian theory would argue that for parties to maximize their appeal to the electorate, they must closely respond to changing value preferences of this electorate. (Note that for such a utilitarian theory, the changing value preference of the electorate is a given, rather than a matter to be explained.) Finally, a materialist theory would argue the priority of social change (without necessarily explaining such social change), so that in this case the similarity in changing concerns is caused by one underlying process of social change.

Certain of these positions cannot be accepted or rejected using any empirical data; they are non-disprovable. We can, however, note the close correspondence between long-term trends of moving averages and one economic indicator.

Figure 4 depicts the long-term trend of the social indicator, percentage unemployed of the total labor force from 1912 to 1956, along with the trends of both party platforms' concern with *Wealth-other* for the same period.[4] In the long run, a change in the social indicator is correlated with changes in the Democratic platforms .94, and with changes in Republican platforms .91.[5]

[4] It was necessary to use moving averages rather than linear regressions to estimate these trends, since unemployment data are unavailable for years before 1900 and since that time the relationships are better described by parabolas than by linear regressions. The unemployment five-entry moving average was determined by averaging unemployment for five four-year periods ending in election years — moving by dropping the first four years and adding an additional four year period, e.g. 1901-1920 = first average; 1905-1924 = second average.

[5] Correlations of *Wealth-other* trends with other economic indicators were much smaller than those with unemployment.

		Economic Indicator	
		Wholesale Price Rate	Business Failure Rate
Wealth-other Trend	Democratic	–.26	.08
	Republican	–.43	.33

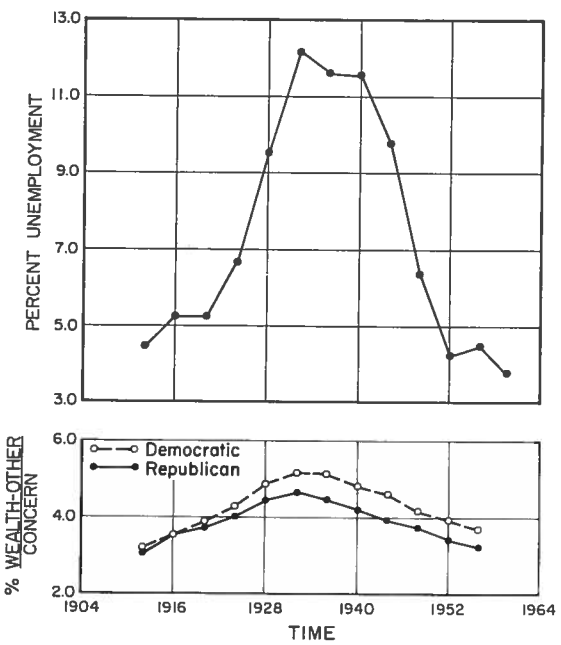

Fig. 4. Twenty-year moving averages of unemployment and Democratic and Republican *Wealth-other* concern, 1912 to 1956.

From these trends, however, it is not obvious whether long-term changes in the economic performance of American society precede, coincide with, or follow long-term changes in party platform concern with *Wealth-other*. To shed further light on this important issue, we compared the long-term trends in the party platforms' concern with *Wealth-other* with the long-term trends in unemployment for the year preceding the election, the year of the election, and the year following the election.

Table I

Correlations of long-term trends in unemployment with Wealth-other concern (N = 12)

		Unemployment Rate		
		Year Preceding Election	Year of Election	Year Following Election
Wealth-other Scores	Democratic	.93	.88	.81
	Republican	.90	.90	.88

In spite of the small number of observations, some generalization does seem warranted. In the first place, value change does not precede economic performance, as can be seen in the smaller correlations in the third column of Table I. Secondly, in the case of the Democratic platforms, the long-term trend of unemployment in years preceding the elections predicts long-term trends in *Wealth-other* concern somewhat better (it explains 10% more of the variance) than do long-term trends of

unemployment in election years themselves. The Republican party platforms *seem* to relate to unemployment in a different manner. Inspection of the scatter diagrams suggests, however, the relationship between long-term unemployment and *Wealth-other* trends is more curvilinear in the Republican than in the Democratic case. The linear correlation coefficients become thus an inappropriate measure hiding the true relationship and the consequent similiarities between the parties in this regard. The absolute values of *Wealth-other* residuals better indicate these similarities.

TABLE II

*Sums of Wealth-other residuals (absolute values) for periods of low unemployment**

		Unemployment		
		Year Preceding Election	Year of Election	Year Following Election
Wealth-other Scores	Democratic	1.3170	1.8629	1.9032
	Republican	0.9774	1.2503	1.2943

* Periods of low unemployment are defined as those years in which the enemployment rate falls below the median for all the years under consideration.

For both Democratic and Republican trends, in periods of low unemployment there is an increasing spread about the regression line from years preceding, and years of, to years following the elections. Why would this be the case? It would seem that in times of economic prosperity, unemployment in the year of the election is less noticeable, the concern with the economy less urgent and the estimation of the economic situation rather insecure (because the year is far from completed). Consequently, the perceptions of the state of the economy are determined more by the completed year preceding the election. On the other hand, in times of high unemployment, concern with wealth and the economy in general is so urgent and the failing state of the economy so obvious that the unemployment rates of the years preceding the elections and the election years themselves equally well predict the trend of concern with *Wealth-other*.

The fact that the relationship between long-term unemployment data and concern with *Wealth-other* is more curvilinear in the Republican than in the Democratic case warrants one further remark. This difference between the two parties indicates that, in the long run, during periods of high unemployment, the increase in unemployment has a smaller effect on *Wealth-other* concern in Republican than in Democratic party platforms. In other words, in times of high unemployment, the Republican party is less sensitive to changes in unemployment than is the Democratic party. At any rate, to explain all these covariations the discussion has relied on a theory that long-term changes in economic performance cause long-term changes in value concern with *Wealth-other*.

SHORT-TERM CHANGES

Having discussed some internal and external dynamics of long-term trends, I shall now turn my attention to the deviations from these long-term trends. The long-term trend indicates the expected level of concern for a particular party and a particular time. Deviations (residuals) from this standard therefore indicate short-term movements which must be attributed to special circumstances of the political process, i.e., the internal dynamics of particular elections: ahistorical issues, personality conflicts and the like.

Plotting the Democratic and Republican residuals from the long-term linear trends,[6] Figure 5 further illustrates the shortcomings of the linear regression as an estimate of the long-term trend: the deviations are generally negative in the early and late periods and positive in the middle period, indicating that residuals from a linear regression are still affected by a long-term historical trend.

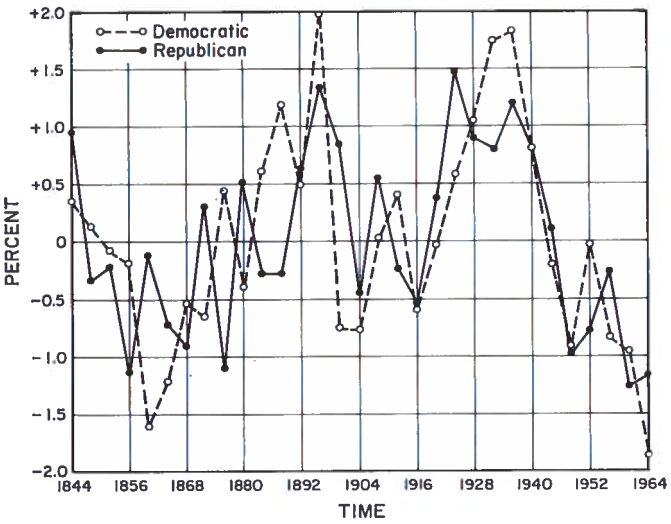

Fig. 5. Democratic and Republican *Wealth-other* residuals from linear regressions of raw scores.

For this reason, I prefer to use deviations from the moving averages (as described by Figure 6) for the analysis of residuals.[7]

[6] $R_i = Y_i - \hat{Y}$ where R_i = residual for a particular year, Y_i = *Wealth-other* concern for that year, and $\hat{Y} = a + bx_i$.

[7] $R_i = Y_i - \hat{Y}_{i-2 \text{ to } i+2}$, where R_i = residual for a particular year, Y_i = *Wealth-other* concern for that year, and $\hat{Y}_{i-2 \text{ to } i+2}$ =

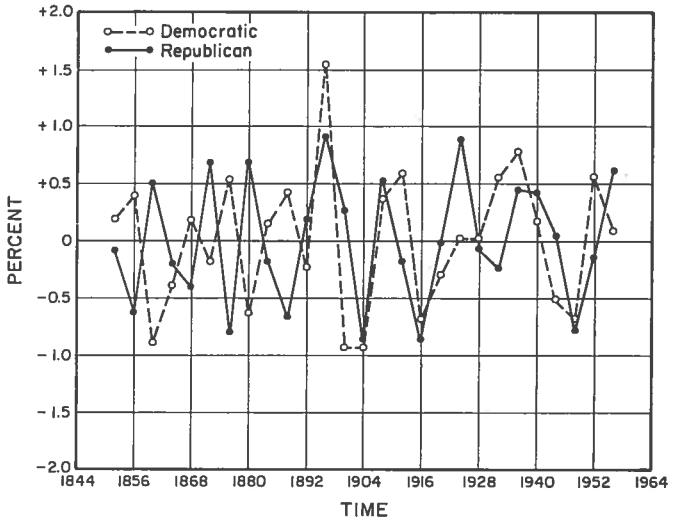

Fig. 6. Democratic and Republican *Wealth-other* residuals from moving averages.

Though at first sight the observations look rather random, closer inspection reveals a most interesting change over time in the nature of the relationship between Democratic and Republican campaign-to-campaign deviations. Whereas in the earlier periods the residuals are either negatively correlated or completely unrelated, in later periods they seem to be positively correlated.

In order to demonstrate this change in the relationship, I correlated the residuals of the two parties for the first ten election years, then for the second through eleventh, and so on, producing a ten-entry moving correlation (presented in Figure 7). Although each of these correlations

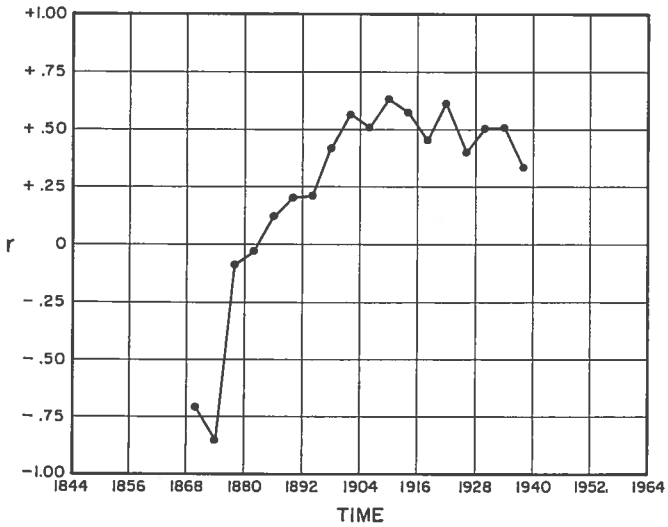

Fig. 7. Ten entry moving correlation of Democratic and Republican *Wealth-other* residuals.

moving average whose middle entry is *Wealth-other* concern in year i.

uses only ten entries, the trend from correlation to correlation is consistent enough to permit some further generalizations. Until about 1892 there appears a strong negative correlation between the two parties: if one party increases its concern with *Wealth-other*, the other party is likely to have decreased its concern with the category. If a common external variable is at all responsible for the pattern of these deviations, we must assume that each of the parties responded to that variable in an opposite manner. A more plausible explanation of this finding is that in the absence of information about the values of the electorate and changes thereof, the Democratic and Republican parties depended on reacting to each other in their competition for the electorate, so that concern with the value *Wealth-other* was closely related to the dynamics of inter-pary conflict. This inter-party conflict, as a main determinant of short-term fluctuation in value pronouncements, was further intensified by the deep political and regional cleavages of the latter half of the nineteenth century: Slavery, the Civil War and Reconstruction. Political parties are unlikely to respond in similar fashion to electoral *Wealth-other* concern when they are so deeply divided on so many other issues. Once these overbearing domestic issues had been settled, the residuals became more positively correlated, for the campaign-to-campaign responses became more dependent on common external factors. (Namenwirth and Lasswell, forthcoming.)

To demonstrate the relation of campaign-to-campaign deviations to these external factors in the late period, I correlated the residuals with a number of social indicators. Using election results as a political indicator,[8] I found that the correlation between the percentage of the popular vote for the Democratic candidate and Democratic *Wealth-other* residuals is .41, while the same correlation for Republican data is .16. The immediate interpretation would be that for the Democrats to talk about *Wealth-other* in excess of the expectation for the election year leads to a moderate increase in their percentage of the vote, while for the Republicans, such a response has little or no effect on election outcome. If indeed this relationship were true, a utilitarian theory could explain the increasing long-term differentiation in concern with *Wealth-other* between the two parties. However, the short-term relationships between residual *Wealth-other* concern and Democratic share of the vote are largely spurious.

[8] Since this study deals only with the platforms of the two major parties, I calculated percentage share of the vote using total votes cast for the two major parties as the base. E.g. Democratic percentage = number of votes for Democratic candidate divided by the total number of votes cast for both the Democratic and Republican candidates. Correlations obtained using election data computed by other methods (percentage of total electoral vote) were not so large.

Let me explain. In the later period, during times of high unemployment, there was a greater likelihood that Democrats would win the presidency. The correlation of unemployment rate residuals with Democratic percent of the vote is .50, and with Republican percent of the vote, –.50. Furthermore, election year unemployment residuals correlate with Democratic *Wealth-other* residuals .56, and only .01 with those of the Republicans. Figure 8 represents these triadic relationships in a causal model according to Blalock (1960; 1961). In the case of

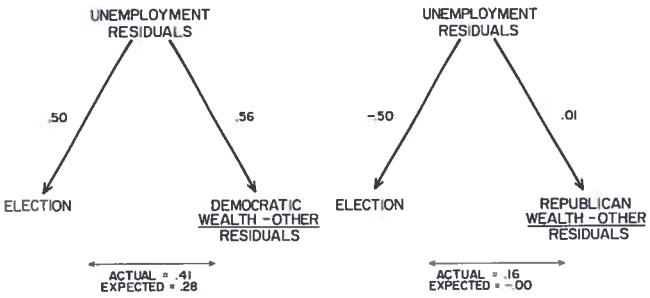

Fig. 8. A causal model.

both platforms, the conclusion seems clear: variation in unemployment causes for both parties a change in election results, and in the case of the Democrats, a campaign-to-campaign fluctuation in the concern with *Wealth-other*.

Causal analysis using contemporaneous covariation cannot settle the temporal priority of one variable over another. To bolster the causal conclusions, I examined the relationship while lagging and leading unemployment data. The results of this analysis are presented in the following table.

TABLE III

Correlations of unemployment rate residuals with Wealth-other residuals

		Unemployment		
		Year Preceding Election	Year of Election	Year Following Election
Wealth-other Residuals	Democratic	.46	.56	.27
	Republican	–.05	.01	–.19

This table indicates that for the Democratic platforms, change in unemployment precedes rather than follows short-term change in concern with *Wealth-other*. However, this time lag must be very small indeed, since unemployment in election year is more strongly associated with *Wealth-other* residuals than is unemployment in the years preceding elections. Republican platforms, on the other hand, are completely insensitive to short-term changes in unemployment. Considering the fact that the Democrats are more dependent on the good will of the workers than are the Republicans, their greater sensitivity to unemployment is not surprising.

The question therefore arises: are the Republican party platforms more sensitive to economic changes directly affecting their constituencies, e.g. business failure rates and indices of inflation?[9] As Table IV indicates, both parties are sensitive to the selected indices. Considering

TABLE IV

Correlations of economic indicator first order differences with Wealth-other residuals

		First Order Differences		
		Wholesale Price Index	Consumer Price Index	Business Failure Rate
Wealth-other Residuals	Democratic	–.56	–.44	.35
	Republican	–.45	–.20	.30

the small differences in these correlations, overinterpretation is an obvious danger. Nonetheless, there is a clear trend: the Democrats are more sensitive than the Republicans in al cases. Also, the differences between Republicans and Democrats are somewhat larger in regard to the unemployment indicator than in regard to the indicators of inflation and business fortune and risk. Consequently, the analysis of short-term trends provides information about both the internal and external dynamics of these changes — which are in most cases at variance with the dynamics of long-term trends.

DISCUSSION

My paper is a clear example of inference based on content analysis, and let us face the problem of inference without further delay. Inference is but an elegant and fashionable term for after-the-fact induction (Namenwirth and Lasswell, forthcoming; cf. Kyberg, Jr. and Nagel, 1963).[10]

[9] In order to isolate short-term changes in these indicators, I used first order differences (e.g. election year business failure rate minus the business failure rate for the year preceding the election). I felt that, in the case of inflation and business failure, the relative change from year to year would be a more pertinent indicator of the perceived state of the economy.

[10] Unfortunately, the concept of inference is often used rather loosely in content analysis. Whereas, when speaking of inference, the content analyst refers to the problems of induction and inductive reasoning, the more frequent connotation of inference is one of deductive reasoning. The real issue, however, is of different order. Many scholars in the humanist tradition contend that inducing lawlike regularities from linguistic data creates a special class of problems over and above the already formidable problems of induction in general. The arguments for this contention seem spurious.

Fishing is the more derogatory term, the sole difference being that in the case of inference, the fisherman not only counts his catch, but also tries to account for it. To put it on a somewhat more formal basis, the rules of induction are either unknown, unspecified or adapted to the findings at hand. It is this circular element in much induction which is especially bothersome in content analysis. It also explains the curious asymmetry of our procedures. It is relatively simple to infer from content the causal predispositions of the communicator — attitudes, values, drives and so on; but by knowing the causal factors, it is hardly possible to predict the ensuing communications. What is clearly needed is a causal theory of communication, of symbol construction, and of language behavior.

My own inquiry is subject to the above limitations. However, analysis of *Wealth-other* concerns in party platforms does produce some interesting similarities and differences between the Democratic and Republican value articulations. In the long run, both parties move closely together, although the Democrats increase their concern at a greater rate than do the Republicans. Consequently, the discrepancy in these value articulations has tended to increase in the past 120 years. Still, in these long-term movements, the parties respond in a similar manner to the overall changes in the economic environment.

The short-term, campaign-to-campaign changes are quite another story. In the first place, the relationship of value articulations between the two parties changes over time. Whereas in the earlier period (1844-1896) the parties move in opposite directions from campaign to campaign, in the later period (1896-1964) they tend to move in similar directions. Secondly, while the Democratic party responds immediately to short-term changes in the economy, the Republican party does not respond at all to unemployment and only weakly to other indicators of the economic cycle.

The present analysis confirms many well-known historical generalizations. It demonstrates that since the middle of the last century, the Democratic party has become the progressive party, while the Republicans have increasingly lagged behind. It confirms the greater sensitivity of the Democratic party to the fate of the economy, to the plight of the working man, as well as to the changing fortunes of the well-to-do. Since 1860, the Democrats have consistently shown greater concern for wealth than the Republicans; they seem to be more in tune with the materialism so characteristic of our industrialized society. The Republican party, on the other hand, is either not sensitive at all, or far less sensitive than the Democratic party to economic change, particularly in regard to short-term changes. This indifference to changing social circumstances can well be explained in terms of political philosophy: it is the earmark of conservatism. Conservative value articulations remain directed towards the preservation of the past conceptions of historical reality and ideological reconstructions.

The preceding interpretations stand in need of further confirmation. After all, these observations are based on the internal and external dynamics of only one value subcategory. The projected analysis of all the other categories in the Value Dictionary may well lead to further specification, modification, if not outright rejection of the present interpretations. This is especially so in regard to propositions about the major theoretical issue. Yet in respect to certain wealth values, the conclusion is obvious. In the total scheme of values, social changes cause rather than follow the changing priority in wealth values, at least since the beginning of this century. Whether the same is true for all values, and if so, whether it is true to a similar extent, is an empirical question which Harold Lasswell and I hope to answer in our next book. The present exercise has been useful in sharpening the tools of analysis, measurement procedures as well as conceptual instruments, for this more encompassing endeavor.

REFERENCES

Blalock, Hubert M.
1960 *Social Statistics* (New York)
1961 *Causal Inferences in Nonexperimental Research* (Chapel Hill, North Carolina)

Democratic Platform Committee
1964 *One Nation, One People*

Kyberg, Henry E. Jr., and Ernest Nagel (eds.)
1963 *Induction: Some Current Issues* (Middletown, Connecticut, Wesleyan University Press)

Lasswell, Harold D., and Abraham Kaplan
1950 *Power and Society* (New Haven, Yale University Press)

Namenwirth, J. Zvi, and Harold D. Lasswell
Forthcoming "Culture vs. Social Action in the Explanation of Social Change. Changing Language in American Party Platforms: A Computer Analysis of Political Values"

National Industrial Conference Board
1964 *Economic Almanac, 1964*

Porter, Kirk H., and Donald B. Johnson
1961 *National Party Platforms 1840-1960*

Stone, Philip J., *et al.*
1960 *The General Inquirer: A Computer Approach to Content Analysis* (Cambridge, M.I.T. Press)

The New York Times
July 12, 1964

The World Almanac

U.S. Bureau of Census
1960 *Historical Statistics of the United States, Colonial Times to 1957* (Washington)
1965 *Historical Statistics of the United States, Continuations to 1962 and Revisions* (Washington)

A Computer Systems Approach Towards the Recognition and Analysis of Content*

HOWARD P. IKER and NORMAN I. HARWAY
Department of Psychiatry, School of Medicine and Dentistry
University of Rochester, Rochester, N.Y.

This paper explores a method of content analysis which allows the user to discover what his data are about without having to furnish a priori *categorizations within which to classify these data. Based originally on problems encountered in psychotherapy — the question of the relationship between the substance and structure of oral communications between psychotherapist and patient — the paper also explores the possibility of adapting this method of content analysis to other kinds of data. The initial task of this analysis is to* generate *informational categories with which many content analytic systems begin. Based on an associational approach, the basic unit of information is the word itself. A matrix of intercorrelations among words is eventually factor analyzed to determine systematically the common factors which may account for the matrix in a meaningful way. To reduce the number of words to a manageable size, a system of programs (WORDS) has been developed. This paper then considers the current use of the WORDS System, its goals and structure, the results and implications of some current research, and plans for the future.*

The key question which we have been exploring for almost six years (2, 3, 4) is whether there exists a method for content analysis which will allow the user to discover what his data are about without having to furnish *a priori* categorizations within which to classify these data. Any other currently used content-analytic system of which we are aware requires, at the least, that the user furnish a set of categories to which the various portions of the text are to be allocated. It is our purpose to demonstrate that there is an alternative to this approach.

Our original interest in the area was in the process of psychotherapy, in the changes in cognitive organization which occur as a result of treatment, and in the process of change. Psychotherapy, if we exclude for the moment certain of the behavioral therapies, involves the user of oral communication to modify, among other things, a person's perspective of himself and of his world. To the extent that this occurs, this change should be reflected in the individual's verbal behavior. Since the assumption is that the communications of the psychotherapist are influential in accomplishing this change, we are led to the question of the relationship between the substance and the structure of the oral communications among psychotherapist and patient, their change with time and with progress in treatment.

Over the years during which we have been developing the system for making such analyses, it has become clear to us that the system can be applied to kinds of data other than those embodied in psychotherapy materials.

The basic assumption upon which this system rests is that there exists sufficient meaning within the word and within the temporal associations among and between words to allow the elicitation of major content materials and categories. In short, it is the initial task of our approach to *generate* the kinds of informational categories with which many content analytic systems begin.

Utilizing an associational approach as the foundation for our method, we take, as our unit of information, the word itself. Dividing an input document into segments of time, or segments of equal length, or equal numbers of sentences, etc., it is possible to count the frequency with which each word occurs in each such segment. Using these data, intercorrelations among all words may be obtained; operationally, these intercorrelations represent the degree of co-occurrence, i.e., association, between words as they are observed across successive units of the data-base. This matrix of intercorrelations may then be factor analyzed to determine, in a systematic fashion, if

* This is a draft of a paper commissioned by and prepared for the National Conference on Content Analysis, November 16-18, 1967 at the Annenberg School of Communications, University of Pennsylvania, Philadelphia, Pa. The research was supported in part by U.S. Public Health Service Grant MH-10444, National Institute of Mental Health. We wish to acknowledge also the aid of Miss Janet Barber, Mr. Gerald Leibowitz, and Dr. Edward Ware.

there are common factors which can account for the obtained associational matrix in an efficient and meaningful way.

In a factor-analytic approach towards this kind of data, it is necessary to reduce the number of different words that are examined. This criterion, a function of machine-size, program running time, and factor interpretability demands that the number of variables (words) analyzed be held to some reasonable minimum. It is the job of the system of programs which we have developed, WORDS, to allow this reduction and the subsequent statistical analyses that are necessary.

This paper will discuss the current implementation of the WORDS System, its goals and structure, the results and implications of some current research, and will then discuss our plans for the future concerning changes in the system and directions for future research.

IMPLEMENTATION OF THE WORDS SYSTEM

Computing Facility

WORDS has been developed at the University of Rochester Computing Center. Until recently, all large scale work in this facility has been run on an IBM 7074 computer. The 7074, a second generation, medium-speed machine, has characteristics including extremely powerful input/output (I/O) hardware and logic, and highly flexible scatter read/write commands. It is a fixed word length machine (five characters or ten digits with sign per machine word) with decimal arithmetic and hardware supported floating point operations. The configuration at the University of Rochester operates with a 10K core and is supported by eight 729-IV tape drives and a 1301 disk file. Final output from the system must be to tape with all printing and punching done offline on an IBM 1401 configuration.

This configuration runs under control of the Computing Center's resident monitor which operates the system on a batched queue basis. As such, all daytime runs operate under closed-shop conditions with all operations handled by a computing-center machine operator.

System Goals

Before beginning actual programming on WORDS, a set of system goals was developed. The considerations dictating choice of these goals were based on the several uses to which WORDS could be put. WORDS had to be capable of large scale, repetitive, methodological investigations; as such program running times had to be as efficient as possible. WORDS had to be useful as a production device; as such, ease of system use and good turnaround time were needed. Finally, WORDS would probably be used by other members of the University and therefore should not be difficult to learn.

With these considerations, as well as those dictated by the method itself, a set of generalized criteria were developed to help in directing the systems work and programming applications to follow.[1]

User-Orientation

WORDS was designed to be as easy to use and to learn as possible. Despite this desire, the final results are far short of the mark as can be attested to by a 110 page user's manual. While the system is not inordinately difficult to use and does not require any extensive experience with computers or programming, its use clearly requires some amount of training. Considering the complexity of the total system itself, this requirement is not unreasonable; nevertheless, both we and the users would be happier were the method easier to understand. Future implementations will result in a much more easily used system. This decrease in difficulty will come about for two reasons; the first is a direct result of our experience with the current system and knowledge of how we could make things easier even within the current system itself; the second reason stems from the greatly increased sophistication and flexibility of third-generation software which substitutes the operating system of the computer for much programming effort on the part of the developer.

Because WORDS does not demand any computer experience of the user it must be able to run under control of the target computer's resident monitor in a closed-shop environment. Since WORDS contains a large number of programs any, or all, of which can be called in any order appropriate to the purpose of the run, WORDS itself must be monitored. For a second generation resident monitor such a consideration demands a systems monitor capable of operating with and communicating with the resident.

Flexibility

WORDS is designed so that the user should have no great

[1] Throughout the development and programming of the WORDS System, the advice and assistance of Mrs. Barbara Rothe has been invaluable.

difficulty in manipulating his data as is necessary for production of appropriate output. Thus, WORDS has been written to allow the user to configure a run with as many or as few programs as necessary, to utilize, wherever possible, mnemonics rather than numeric information for control purposes, to afford as effortless a handling of I/O as possible, and to allow a wide range of output formats.

Efficiency

Typical data-bases in our current research involve files of approximately 25,000 words. Typical run configurations often run more than 20 separate program calls for successive manipulations of this data. A total run, then, can easily involve the computer in the manipulation of well over one million records. On a machine with the speed of the 7074, then, a high degree of program efficiency is a clear necessity.

Protectiveness

Use of WORDS almost invariably demands repetitive runs on the computer with the input for any given run being based, at least in part, on output from prior runs; in such a case, some devices are needed to protect prior data from possible destruction during a run. Additionally, a complete run configuration for any extensive processing by WORDS can generate a complicated calling sequence with massive I/O operations so that a software or hardware failure is always a possibility. WORDS is designed to fail-safe and to produce as much diagnostic information for the user as possible.

Appendix 1 may be consulted for further information on the system. It details the structure of the WORDS System: its systems organization, data organization, program organization and a list and description of each of the major programs in the system.

CURRENT RESEARCH

Methodologic Issues

The UHH Approach

Over the past several years, apart from the complete reprogramming of WORDS, our major research emphasis has been to investigate some of the methodological problems confronting this technique. One of the most important of these issues, both on a practical and theoretical level, derives from the technique used for the reduction of the number of different words found in the raw data base. The question which we have investigated was whether alternative methods for reduction of these words could be found that did not involve the extensive use of synonymization.

In our analysis of a typical data base — five psychotherapy interviews — we encounter about 1300 different words making up the total 25,000 word protocol. Since the maximum matrix with which we can work is 215 variables, we must make reductions comprising 83% of the total number of different words.

The method we initially employed was based on a four phase process. Utilizing the PARSE programs, all articles, prepositions, conjunctions, etc., were removed. The next phase used STRIP to change all words into root form. The third, heavily employing synonymization, combined words having the same basic meaning and being used in the same fashion. Lastly, a list of remaining different words, sorted by frequency of occurrence, was generated; beginning with the word of highest frequency, this list was downcounted to reach 215 different words which were then subjected to matrix analysis. It is the synonymization phase which we have found most difficult in implementation and most dangerous in terms of objectivity. Synonymization requires two basic commitments from the user: the first being extensive amounts of time and the second being the use of subjective and potentially unreliable judgments. In many respects, synonymization places the same demands on the user as does the typical content-analytic method of *a priori* categorization. Except for the fact that the data itself, rather than the initial interests of the investigator, generates the potential categories (generic word), we are faced in synonymization with the same degree of subjectivity and inefficiency as we would be were we to have begun word reduction with a set of categories into which the data was to be allocated.

With a synonymizing procedure, the time demands placed upon us in reduction of the data in a typical set of five interviews was considerable. At least a month was required, from both investigators, with a heavy investment of twenty-seven pre-factor runs on WORDS being necessary before producing the final rotated factor structure representing the content-recognition portion of the analysis.

More important than time, however, was the constant requirement that we exercise our own judgment as to when two words were being used in a fashion and with a meaning such that they could be combined into one. While WORDS is implemented to make this task easier and more reliable

than before (cf. Appendix 1: HSTRY, IDIOM, TEXT 1 and TEXT 2, PARSE, etc.), there is little question but that our ability to maintain a high degree of freedom from *a priori* needs and ideas in the selection of combinations decreased as the time wore on and the combinations became more and more difficult to locate.

Our experience, then, in the analysis of data using *ad hoc* synonymization procedures to carry a large bulk of the reduction process demanded that we investigate other techniques that were more efficient and less subjective. Accordingly, we began work with a different approach in which synonymization played a minimal role and in which major reductions were accomplished by deletion procedures.

This new technique, which we have euphemistically labeled UHH (Untouched by Human Hands) as opposed to SYN (Synonymization), operates according to a generalized set of rules. The application of these rules, rather than *ad hoc* decisions deriving from our own inspection of the data, clearly reduces both the time required and the subjectivity involved in the reduction process.

UHH rules currently fall into two phases: pre- and post-factoring. In the pre-factor phase, we first parse and then subject the data to a common STRIP run for the purposes of de-inflection and a change to root form of all comparatives. Following this, EDIT is applied. EDIT is used to make four kinds of change: (1) deletion by part of speech so that all articles, prepositions, conjunctions, etc., are removed; (2) deletion of certain words which carry very little meaning outside of context, e.g., *sort*, *still*, *be*, *thing*, *ago*, etc.; (3) deletion by the combination of word/speech categories so that, for example, *kind* (adjective) is retained while *kind* (noun) is deleted, or *like* (verb) is retained while all other forms of the word are dropped; (4) lastly, a low level of *pre-determined* synonymization is applied in which generic words are created to subsume a set of other highly related high frequency words, e.g., NO is held as a generic word which will contain all occurrences of *neither*, *never*, *nobody*, *none*, *nor*, *not*, *nothing*, and *nowhere*.

It is of importance to the UHH approach to note that this kind of synonymization rule is applied to data prior to any analysis of the data and, where feasible, is applied prior to any inspection of the data itself.

After subjecting the data to this set of pre-factor rules, a downcount is taken on a frequency ordered list and the 215 words with the highest frequencies are then subjected to factoring. Post-factoring rules serve a common goal: the improvement of the obtained factor structure. Such improvement can come about in two major ways within the constraints of our methods and techniques; the first kind of improvement obtains factors which have a better statistical structure such that loadings are improved, amount of variance extracted is increased, and factor independence is better. The second kind of improvement, obviously not independent of the first, is to improve the "meaningfulness" of the factors.

In our attempt to improve both structure and content, we have begun to investigate post-factor rules for the deletion and/or synonymization of words. There are basically two such rules. The first derives from clusters within the factors themselves. Thus, one factoring run yielded a factor with the days of the week heavily loaded within it; we utilized this information to produce a generic "TIME" term subsuming the days and thus opened the matrix size for the inclusion of seven additional words should this be indicated.

The second rule which we make use of in post-factoring derives from the fact that common usage holds loadings less than .30 as being fundamentally uninterpretable. Thus, we are also investigating the results from re-factoring after having dropped all words which never obtain a loading greater than .29 anywhere in the obtained factor structure. Obviously, both of these rules result in a reduction of total matrix size; one of the questions, within a UHH approach, with which we are concerned is whether we better obtain our goal of improved structure by re-factoring with a smaller size matrix or whether it is more fruitful to include additional words (formerly unincludable) now available because of the open slots in the matrix. Our initial results suggest that replacements, rather than a size reduction per se, is the more appropriate technique.

Using a UHH approach, the time for complete analysis of the same set of data mentioned earlier, has changed to approximately four hours of investigators' time as opposed to almost a month using SYN and, as would be expected, the number of computer runs has sharply reduced. Typically, four runs carry us through initial screening, parsing, deinflection, editing, factoring and re-factoring. Depending upon turnaround times and daily load requirements at the university computing-center, we can reasonably expect to complete analyses in something less than a normal work week. With SYN, turnaround time and repetitive runs, made six weeks a minimum.

The results, with UHH, have been quite encouraging. Before beginning the UHH factoring of interview 23-27 of subject Pl. we had available the SYN results on that same data.[2] Accordingly, we used, as a partial criterion, the

[2] This data consists of a set of 462 consecutive psychoanalytic treatment sessions wich were tape recorded and made available to us by Dr. F. Gordon Pleune. We are grateful for his help and his cooperation in the analysis of this data.

comparability of the two factor structures to make some judgment about the viability of a UHH approach. While the factor structures were not identical, there was sufficient similarity between the two to make us believe that the method should be refined further.

Table 1 illustrates the kind of structure and the basic similarity between the two approaches.[3]

TABLE 1

Part 1. *Comparison of SYN and UHH Factors**

(Varimax rotated loadings truncated at < .30)

SYN FACTOR 3		UHH FACTOR 12	
Old	95	*Old*	94
Change	94	*Clothes*	91
Dress	85	*High*	91
Look	83	*School*	83
Friend	79	*Friend*	82
Okay	77	*Dress*	62
School	77	*Look*	53
Clothes	75	*Enjoy*	50
Sloven	72	*Definite*	47
Attract	70	*See*	45
Relax	63	*Long*	43
Keep	56	*Apologize*	41
Apologize	41	*Always*	34
Good	38	*Keep*	34
Shave	38	*Lot*	34
Meet	36	*Attention*	32
Differ	34	*Normal*	30
Peculiar	33		
Attention	32		
See	32		

* SYN refers to word reductions via synonymization as well as deletion; UHH data is reduced without synonymization.

Part 2. *SYN Factor 3 High Scoring Segment*

Friend of mine and I and every time I see this old high school friend, I am always dressed in old clothes. And one day, I sort of apologized for always seeing him this way and he said he never saw anybody look as good or as well in old clothes as I do. I don't know if I do it for, I know that I have done this, that I have gone to parties not shaven and dressed this way and would go weeks in high school, not weeks but a whole week, with dressing this way or slovenly. But I, I guess it, it would attract attention. I know if I see somebody dressed this way and unshaven for a whole week I would look at them myself.

[3] In this, and in all other analyses reported in this paper, a combination factor of 5 has been used in preparing the data for correlation. Thus, if psychotherapy data is being analyzed each successive set of five minutes is combined and analyzed as one observation; analogously, in analysis of the book data to be reported later, each successive set of five pages has been combined and analyzed as one observation.

Factor Scoring

One of the more important uses for WORDS, in a post-factor environment, is for content analysis. It is of importance to us that we be able to investigate content changes over time, across speakers, under different circumstances, etc. Our first step for implementing this goal, involves locating those portions of the data base from which specific content and/or thematic areas were being elicited. The SCORE program in WORDS is designed to accomplish this task.

With the final factor structure loadings as the prime data, SCORE will scan the data base, from which these loadings were obtained, and will assign to each of the observations (segments or combinations thereof) a numeric factor score. Using only factor loadings greater than .49, this score is obtained, for each factor, by using the loading of all words on that factor, as a multiplier for the frequency of occurrence of that word in the observation being scored; these separate word-scores are then summed over all the selected words on that factor yielding a factor score. In order to afford comparability between factor scores, SCORE computes standard- as well as raw-scores for each factor on each observation.

Table 1 also illustrates the use of SCORE. The segment presented in Table 1 is the highest scoring segment for the SYN factor seen in that table. This is a typical result and there seems little question that SCORE can quite well locate that portion of the data base which is heavily saturated with the material that is helping to elicit the factor which is being scored and that the material located is consonant with the content of the factor.

Illustrative Results

As part of our continuing research program, we have recently begun application of the system to data other than the series of over four-hundred continuous psychotherapeutic interviews which comprised the initial data base from which the system was developed. As a suitable vehicle, we chose a set of two psychotherapy interviews recorded over fifteen years ago (1). Our choice of these two interviews was dictated by the fact that they form the nucleus of a book, *Comparative Psycholinguistic Analysis of Two Psychotherapeutic Interviews*, which was edited by Gottschalk and published in 1961: the purpose of the symposium, from which this book was generated, was to bring together several workers in the area of content analysis and to bring their different approaches and skills to bear on the same set of two interviews. The results of

these different analyses form the major part of the book; our hope was that analysis of these same two interviews by WORDS would yield data which would illustrate the utility of the system by allowing comparison of our results with those of some of the members of the symposium. Our purpose in presenting these results is not to offer or interpret further information as regards the particular case under analysis; rather, by showing some portion of our results we aim only towards demonstrating the utility of WORDS as a method for content analysis.

The data incorporated in the two interviews analyzed in the symposium is based upon two separate meetings, interviews number 8 and 18 of the patient under treatment. Following a description of the patient and the comments of the psychotherapist as to the content of the interviews, the book then details the interviews themselves, a set of physiological observations as to skin temperature and heart rate for both therapist and patient (on a minute-by-minute basis) and then presents papers by Strupp, Jaffe, Mahl, Gottschalk, et al., and DiMascio. The papers by the last four authors present materials which are reported in tabular or graphic form in the text in a quantifiable fashion; thus, Jaffe uses the verbal diversification index (type-token ratio) and a percent present-tense index, Mahl defines a speech disturbance ratio and silence quotient, Gottschalk presents and scores categories for anxiety, hostility, and schizophrenic disorganization, and DiMascio makes use of the physiological indices listed above.

In analyzing the data of these interviews by WORDS, we used the UHH approach. Following the usual pre-factor rules, we submitted a list of a hundred-and-ninety different words for factoring. With these results as a baseline we then included, as an additional set of twenty-five variables, the various indices derived from the papers presented by the members of the symposium; inclusion was done by representing the variables as though they were words with a frequency equivalent to their "score". Thus, for example, if the patient had had a heart rate of 75 in segment four of interview eight, a variable PHR was included seventy-five times within that same segment; following usual WORDS reduction and summarization procedures PHR would yield a combined frequency of 75 in segment four of interview eight. After analysis of the basic plus auxilliary data matrix, we compared this with the matrix of words only and found little basic difference between the two sets of factors. The substantive data factors were almost the same; that auxilliary data which had been submitted simply loaded *within* various of these factors.

Although we extracted twenty factors, we found to our surprise (perhaps because we were dealing with a data base only 40% of our usual size) that almost 100% of the variance had been extracted by the first nineteen factors.

The results of the analysis have been very provocative. The significant and frequent loadings of so much of the auxilliary data clearly indicates the extent to which content themes extracted by WORDS relate to indices extracted according to other widely different theoretical constructs. It is important, however, to note that some of the relationships are probably artifacts. In Factor 13 we have a loading of .70 for "OUTWARD HOSTILITY" and a dominant loading of .96 for the word *kill*; this simply suggests that the word *kill* plays an important role in scoring the category. On the other hand, such indices as the type-token ratio, physiologic measures such as heart rate and skin temperature, percent of present tense verbs, etc. cannot be reasonably explained in terms of content artifact.

The results of this analysis are too extensive to present in entirety. Rather, we shall show four factors of the second analysis (in which the auxilliary data was included) in order to illustrate the kind of materials produced.

In the beginning of the Gottschalk book, in a chapter by Kanter and DiMascio, a summarization of the content of the two interviews is furnished by the psychotherapist. Table 2 contains a portion of the therapist's description, the significant loadings on Factor 15 of the WORDS analysis, and the segments chosen by the SCORE program as the highest loaded in the data on that particular factor.

TABLE 2

Gottschalk Data Example 1

Part 1. *Therapist statement.* He continued to work at understanding his feelings. In the course of this, he told of blocking himself and hurting himself and handling the humiliation by clowning and playing the buffoon as his father had before him. With a feeling of horror, he told of his identification with being publicly humiliated before those who matter. (1, p. 20)

Part 2. *FACTOR 15.* (Varimax rotated loadings truncated at <.30)

ANXIETY	92	*THERAPIST HEART RATE	−46
HUMILIATE	89	OTHER	44
*EMOTIONAL DISCOMFORT	81	*INTRAPRSNLSCHIZWITHDRWL	−38
*FREE ANXIETY	73	*BLOCKED RELATIONS	−36
SPECIAL	72	*PATIENT SKIN TEMPERATURE	−36
FAIL	64	*TOTAL SCHIZ WITHDRWL	36
UPSET	54	WIFE	35
CERTAIN	50	ALL	−34
WANT	50	EXPECT	−32
*% PRESENT TENSE VERBS	50	MANY	−31
INTELLECTUALIZE	−47	*GRATIFYING RELATIONSHIPS	31
*TYPE TOKEN RATIO	−46		

* Non-verbal categories.

Part 3. *High Scoring Segments 40-41.*

40. being the therapist to this group and then this woman on a program which I identified with. Of course I was really upset for this woman in front of her child. I remember a similar feeling of the greatness of the movie "Bicycle Thief", the humiliation of the father in front of his son when he was caught stealing the bicycle. I thought that was the greatest part of the movie, very much aroused. I am sure I cried at that part of the movie. Uh that really affected me but ah I don't know, and then yesterday I started talking about my father and his humiliation and then his handling the humiliation by clowning

41. and being buffoon and so do I, I clown and play the buffoon. Humiliation is a real issue to me, humiliation in front of people, I think of tarring and feathering someone. I think of it with horror. The humiliation with which they treated collaborators who were stripped and had their heads shaved, especially the woman. It really deeply affects me. I must identify with that — being humiliated ... Gee, I always talk about my anxiety, ha ha, I can talk freely to people, I'm very anxious, I'm very anxious over such and such, almost ready to have them say no, you weren't, or to show them that I wasn't really ...

In a later paper, Mahl presents information as to his analysis of the data using the speech disturbance ratio and silence quotient. Mahl raises a question as to what causes the variations in the speech disturbance ratio; noting that his objective measures are not designed to answer this question, Mahl nevertheless attempts to pinpoint some of the variation by noting that "the therapist's increasingly prodding, insistent questions and comments on the patient's lateness ... are associated with the progressive rise in speech disturbance level ..." Mahl identifies the 29th-32nd minute of interview 18 as being these points. Table 3 presents Factors 11 and 18 which contain the two highest loadings of SDR across all of the extracted factors. Factor scoring for Factor 11 selects segments (minutes) 26-30; scoring on Factor 18 analogously retrieves segments 31-35.

As a final illustration of the results, it is worth considering some of the physiological materials presented and analyzed by DiMascio in another chapter of the book. Because each of the participants in the symposium had utilized his own methods and skills for construction of the various indices to be applied to the data, DiMascio's ability to relate the various physiologic indices to this other data was confined to a series of correlations. What WORDS allows, on the other hand, can be easily seen by inspection of Factor 4 presented in Table 4.

Again, we should note that this re-analysis of the data from the content analysis symposium is not an attempt to offer new information or conclusions about the data analyzed by that group, although it could indeed serve such a purpose. Rather, we cite this information and show these results to begin our attempts in making an assessment

TABLE 3

Gottschalk Data Example 2

(Varimax rotated loadings truncated at < .30)

FACTOR 11		FACTOR 18	
RETALIATE	97	FEAR	78
IDENTIFY	81	*STRUCTURAL SCHIZ WITHDRWL	−75
CONFIDENT	80	MUST	69
REASON	72	YES	58
*INWARD HOSTILITY	71	*SPEECH DIST. RATIO	53
CHANGE	70	NEW	50
YOU (Therapist)	68	RESIST	49
SPEAK	60	COURSE	42
*SPEECH DIST. RATIO	54	MANY	−42
*TYPE TOKEN RATIO	−54	CONCERN	−36
*GRATIFYING RLTNSHPS	53	EXPECT	−36
SEE	47	*TOTAL SCHIZ WITHDRWL	−33
REACT	46	MAYBE	−31
SAY	44	LESS	30
CHANCE	38		
FEEL	33		
TAKE	32		
ACTUAL	31		
TALK	−31		
*FREE ANXIETY	30		
*EMOTIONAL WELL BEING	−30		
*TOTAL SCHIZ WITHDRWL	−30		
High Scoring Segments or Minutes	26—30		31—35

* Non-verbal categories

TABLE 4

Gottschalk Data Example 3

Part 1. *FACTOR 4.* (Varimax rotated loadings truncated at <.30)

FLY	98	HOPE	54
BOY	95	REMEMBER	51
LET	94	*OUTWARD HOSTILITY	47
MUMPS	93	GOOD	44
AGGRESS	91	*INTRAPRSNL SCHIZ WITHDRWL	42
BACK	88	*SELF ESTEEM	40
*THERAPIST HEART RATE	−77	MAYBE	−39
GI	68	PRETTY	39
*OUTWARD HOSTILITY THEME	65	SHOULD	39
FEW	64	*SPEECH DIST. RATIO	−39
*PATIENT HEART RATE	−61	FAMILY	−32
*PATIENT SKIN TEMPERATURE	−58	CAN	31
*INTERPRSNL SCHIZ WITHDRWL	−56	MANY	−30

* Non-verbal categories

Part 2. *High Scoring Segment 52.*

(Prior to this segment, patient describes a private beach used by his family and his discovering a group of soldiers there one day who refused to leave. He speaks of his maneuvers to get them to go, his hopes that they will contract his son's mumps, his wife's fear he will get into a fight; he talks of getting ready to flycast the beach area and remarks that if one is not skillful people behind the caster can sometimes get hurt. He states his inability to tell the soldiers:)

... would you mind moving I'd like to cast this area, I don't let my kids sit there, which I don't. Uh, I just didn't say anything. I let the fly fall a few times thinking I hope one of them gets the fly, and then being afraid though to really hook someone with a fly or whip them with it. Actually, it's a whipping, a good sharp slap. Being afraid I realized well here I'm going through all this indirect aggression and suffering it through and I can't really express it. Even if it were a physical fight I would have been proud of myself to have been able to do it and feeling that I could have carried it off. I was physically in wonderful shape. I'd been swimming and I'm usually in good shape anyway, thinking if it would just help establish a relationship between me and my son. I'd just seen the picture — in which ...

of the validational properties of WORDS and of the factors and factor structures which it produces.

We have been concerned about the validational properties of this method since its very inception. While there are many validational criteria that might be applied, two major aspects of validity have seemed of primary import to us. The first of these two criteria concerns the extent to which the factor "fits" the data. That is, considering each factor independently how well does it identify its portion of the data (as defined by high factor score) and to what degree is that portion of the data selected consonant with the factor. The second validational criterion concerns the factor configuration derived; how well can the data base, or portions of the data base, be described in terms of the factors.

Our attempts to answer these questions with the kinds of data for which the system was originally developed — psychotherapeutic protocols — is difficult. We used the Gottschalk data in our hopes that it would offer enough ancillary information to help us assess these questions. While the data did indeed make it easier for us to assess our factors — by utilizing descriptions of the protocols from the book —it demonstrated that, once again, we would be forced to buttress these factors by judgmental statements made by others as to what was indeed the content of the data base. Since it is precisely in order to avoid such dependence upon judgmental techniques that we developed the system, we felt dissatisfied with the results of the Gottschalk analysis. It seemed clear that what we needed, for analysis, was a data base that had clearly defined content that was known to large numbers of people. In short, we wanted a data base whose content was clear enough and public enough to make a "face validity" approach towards factor assessment a reasonable tenable procedure.

Accordingly, we turned our attention to famous children's books. Such books are relatively short, utilize somewhat restricted language, tend to have clearly defined content, and are usually known, in broad outline, to many people. Having examined a number of possibilities, we selected Frank Baum's *The Wizard of Oz*. This story satisfies all the criteria mentioned above and is certainly one of the most well-known children's books of all time.

In the analysis of *Oz*, we felt that we were posing a fairly stringent but appropriate test for the WORDS System. The major themes of *Oz* are well known both from the book and the movie.[4] Certainly, if the method is viable, one must expect to see content themes and/or materials clustering around the Tin Woodman, Dorothy's flight in the cyclone to the land of Oz from her home in Kansas, the Cowardly Lion, the Scarecrow, the Wicked Witch, the Wizard himself, etc.

Oz was the largest single data base we have ever analyzed, numbering almost 42,000 words. The data was analyzed according to the UHH rules mentioned earlier. After assignment of parts of speech, and deinflection to root form, we were left with a data base of approximately 1400 different words. This list, in frequency order, was down-counted in order to obtain the 215 highest frequency words; these words constituted the UHH analysis. Twenty factors were extracted, rotated by varimax, and factor-scored for each chapter. The factors, presented in Appendix 2, account for 80% of the variance represented by the initial 215 × 215 matrix submitted.

The two questions posed earlier were examined in the light of these results. The first of the questions, the validity of each factor for its eliciting content, may be examined by referring to Table 5. This table describes each of the factors both in terms of its four highest loading words and in terms of a brief description of the content of the total factor. With the presentation of each factor will be found that chapter yielding the highest factor score for the factor and the title of the chapter. Because of the nature of the chapter headings, it can be seen that twelve of the twenty factors can be immediately verified by inspection of the key loadings on the one hand and the chapter title (in which they occur most heavily) on the other; thus, Factor 2 is most heavily loaded into Chapter 5, "The Rescue of the Tin Woodman", Factor 8 in Chapter 14, "The Winged Monkeys", etc. The remaining eight factors do not automatically link with the chapter titles;

[4] *Oz*, now in the public domain, is available in numerous editions. We checked several to confirm the fact that ours was standard in content. Two changes were made to the book in our analysis: (1) Chapter 20, "The Dainty China Country", was omitted since it added many different words for a very small increment in total data; (2) Chapters 23 and 24 were combined since Chapter 24 is but one-half page long. It is worth noting that several differences exist between the movie and the book; we mention the major ones in order not to confuse the reader: The movie has ruby slippers, the book silver shoes; the movie pays much attention to Kansas and reproduces the Kansas characters in Oz while the book does neither; the movie has Oz as a "dream", in the book Oz is "real"; finally, the movie omits any reference to the Kalidahs, the Golden Cap, the Hammer Heads, and the Field Mice.

TABLE 5

Identification of Factors in Oz
(For complete factors cf. Appendix 2)

Factor and four high loaded words	Factor content	Highest loading chapter and title
1. *ax* *oil* *Tin Woodman* *tin*	The Woodman, his body, one time romance with Munchkin girl.	5. "The Rescue of the Tin Woodman"
2. *Oz* (HEAD) *Oz* (LADY) *kill* *send*	Meetings with Oz in his various disguises. Oz's demand that they kill Wicked Witch of West in order to receive their requests.	11. "The Emerald City of Oz"
3. *farmer* *brick* *scarecrow* *road*	The Scarecrow, his creation, stupidity, clumsiness, need for a brain.	3. "How Dorothy Saved the Scarecrow"
4. *Munchkins* *Witch* *East* *woman*	People of the East; freed by Dorothy whose house killed the Witch.	2. "The Council with the Munchkins"
5. *Uncle Henry* *house* *Aunt Em* *bed*	Dorothy's aunt and uncle, her home in Kansas, her trip to the land of Oz.	1. "The Cyclone"
6. *wolf* *lie* *crow* *die*	Attacks by animals on one or more members of group during book; chief such attack is instigated by Wicked Witch.	12. "The Search for the Wicked Witch"
7. *coward* *near* *Cowardly Lion* *heart*	The Cowardly Lion, his attack on Toto, Woodman, Scarecrow.	6. "The Cowardly Lion"
8. *Gaylette* *Quelala* *time* *Winged Monkeys*	The Winged Monkeys, creation of their controlling agent — Golden Cap — by the sorceress Gaylette.	14. "The Winged Monkeys"
9. *Winkies* *tinsmith* *set* *careful*	Winkies, people of West, slaves to Wicked Witch, freed by Dorothy; their special friendship to the Woodman.	13. "The Rescue"
10. *mouse* *Queen Mouse* *safe* *turn*	The field mice; Queen Mouse, saved by Woodman directs mice to save Lion from poppy field.	9. "The Queen of the Field Mice"
11. *spectacles* *Guardian/* *Gates* *want* *Emerald City*	The entry to the Emerald City via the Guardian of the Gates who requires all to wear green spectacles on entrance.	10. "The Guardian of the Gates"
12. *flower* *stork* *sleep* *bright*	The poppy field with its "deadly" fragrance which puts anyone going through to sleep forever.	8. "The Deadly Poppy Field"
13. *terrible* *man* *home* *promise*	The wizard; his trickery of the group; his broken promise to each of them.	15. "The Discovery of Oz the Terrible"
14. *silver shoes* *end* *water* *Wicked Witch*	The magic shoes gained by Dorothy from Witch of East and coveted by Witch of the West leading to her downfall.	12. "The Search for the Wicked Witch"
15. *room* *green* *soldier* *dress*	The palace of Oz in Emerald City where the Group await their audiences with Oz.	11. "The Emerald City of Oz"
16. *balloon* *air* *silk* *make*	The device which first brought Oz to the magic land and which he and Dorothy hope to use to return home again.	17. "How the Balloon was Launched"
17. *pretty* *Hammer Heads* *kind* *dress*	The Hammer Heads, people with projectile heads who bar the group in their journey south to see Good Witch.	22. "The Country of the Quadlings"
18. *pole* *river* *middle* *let*	The ubiquitous pole with which Scarecrow has much trouble. Chief problem: stuck on pole in middle of river.	8. "The Deadly Poppy Field"
19. *tree* *side* *branch* *seem*	The various forests and the scenery encountered by the group in their travels.	19. "Attacked by the Fighting Trees"
20. *courage* *real* *brain* *very*	A generalized factor about the "needs" of the group: courage for the lion, a heart for the woodman, brains for scarecrow.	15. "The Discovery of Oz the Terrible"

however, all relate quite appropriately as can be seen in Table 5 under the "content" which relates the factor content to the chapter content. Thus, Factor 2 describing the various meetings of the group with the disguised Wizard occurs in Chapter 11, "The Emerald City of Oz"; it is, however, in that chapter that all of these particular meetings take place. Likewise, Factor 17 achieves its main loadings in Chapter 22, "The Country of the Quadlings", a chapter almost completely devoted to the group's attempt to reach the country of the Quadlings and their difficulty in following their route since it is barred by the Hammer Heads whom they must circumvent.

Our reaction to these results, then, is that the factors are indeed relevant to the content areas which they identify. This approach has dealt only with the *highest* factor-loading score for each factor; factors, however, have a factor score for *each* of the chapters under analysis and we therefore turned our attention to those areas in which factors were occurring at some significantly high level. The most effective way to present this kind of analysis is on a chapter-by-chapter basis; thus, we are raising the question as to how well the *chapters* are described by the factors as opposed to how well a particular factor dovetails with its highest scored segment. Table 6 presents this chapter-by-chapter analysis. The chapter, its title, and a brief description of its content are furnished; for each chapter, the factor number, standard-score for that factor, and a mnemonic based on the basic factor structure are listed. All factors with a standard score of +2.00 or greater are presented. Factor scores with asterisks indicate the highest score ever obtained by the factor.

Again, the results are very encouraging. In Chapter 1 which is devoted to Dorothy's home in Kansas and her trip, the only factor scoring at 2.00 or more is Factor 5, "Home". In Chapter 2 which details her arrival in Oz, the meeting of the Munchkins, receipt of the Silver Shoes, and her plans to see the Wizard, Factors 4, 14, and 13 score at or above 2.00 thus identifying the "Munchkins", the "Silver Shoes", and the "Wizard".

Close inspection of Table 6 will reveal that certain themes, noted in the "contents" column are not being supported by appropriate factors and that one chapter, Chapter 21, has no factor scored at +2.00 or greater. This chapter is concerned with the cowardly lion's "election" to become king of beasts after he has killed a monster spider terrorizing the other animals in the forest. *Spider* and *monster*, the two words most closely associated with the chapter's theme, did not obtain frequencies high enough to be included in the 215 word list for factoring. Whether enough other words, themselves associated with these two key words are included in the 215 list and potentially available in factors beyond number 20, is presently something we cannot ascertain. The issue of what words are included, and the consequences of missing words, is a topic to which we shall later turn our attention. It is perhaps worth noting that the highest scoring factor in Chapter 21 with a standard-score of +1.29 is Factor 7, "Cowardly Lion".

After concluding this initial analysis of *Oz*, we decided to try another approach which would yield information as to the effects of word choice on factor extraction. In order to do this, we again started with the list of 1400 different words (used in the last analysis to yield the

TABLE 6

Relationship between Factors and Chapters in Oz
(For complete factors cf. Appendix 2)

Chapter and major themes presented	Factor and mnemonic	Z-score
1. "The Cyclone". Aunt Em, Uncle Henry, Dorothy, Toto; trip in cyclone begins.	5. Home	7.81*
2. "The Council with the Munchkins". Arrival in Oz where Dorothy meets Munchkins and Good Witch of the East; learns she has killed Wicked Witch of East and freed Munchkins; gets silver shoes; decides to see Wizard to go back to Kansas.	4. Munchkins 14. Silver Shoes 13. Wizard	5.17* 2.50 2.40
3. "How Dorothy saved the Scarecrow". Begins trip, spends night Munchkin farm, meets Scarecrow, gets him off pole; he joins trip to seek brains.	3. Scarecrow 4. Munchkins 18. Pole	5.61* 3.19 2.33
4. "The Road through the Forest". While walking, Scarecrow tells story of his creation his failure as a scarecrow, his stupidity, etc. They find a cottage to spend the night.	3. Scarecrow 6. Animal Attack 13. Wizard 19. Trees	4.40 3.04 2.34 2.32
5. "The Rescue of the Tin Woodman". Woodman is found, oiled, freed from rust; tells how Wicked Witch of East enchanted his ax making him cut off arms, legs, etc. to prevent marriage to Munchkin girl; joins trip to seek a heart.	1. Tin Woodman 4. Munchkins 13. Wizard 3. Scarecrow 19. Trees	7.33* 4.78 3.31 2.15 2.10
6. "The Cowardly Lion". Lion attacks Woodman and Scarecrow, to no avail, turns to Toto, Dorothy slaps him, cowardice revealed; learns purpose of trip and joins to seek courage from Wizard.	7. Cowardly Lion 20. Needs	6.04* 2.91
7. "The Journey to the Great Oz". Must cross large ditches on back of Lion; attacked by Kalidahs; Lion delays them while Woodman chops down log bridge; escape; come to river and prepare to build a raft to cross.	19. Trees 18. Pole 7. Cowardly Lion	5.73 2.29 2.00
8. "The Deadly Poppy Field". Build raft, begin to cross, Scarecrow stuck on pole middle of river; ashore, others solicit stork to carry Scarecrow back; she does; come to poppies; Lion, Dorothy, Toto succumb; Woodman and Scarecrow carry Dorothy and Toto but cannot move heavy lion.	18. Pole 12. Poppies	8.99* 7.66*

TABLE 6 (Cont.)

Relationship between Factors and Chapters in Oz
(For complete factors cf. Appendix 2)

Chapter and major themes Presented	Factor and mnemonic	Z-score
9. "The Queen of the Field Mice". Woodman saves Queen Mouse from wildcat, she offers help, Woodman makes "truck"; mice drag Lion from poppies.	10. Mice	6.62*
10. "The Guardian of the Gates". Lion awake, trip continues; spend night at cottage discussing hope for success with Oz; reach Emerald City, admitted by Guardian of Gates; all must wear spectacles to prevent blindness from brilliance of city.	11. Emerald City 13. Wizard 20. Needs	8.95* 2.52 2.20
11. "The Emerald City of Oz". Enter city, reach palace, each assigned separate room pending their separate audiences with Oz as a different figure; Oz refuses each pending death of Wicked Witch of West; they agree to kill her.	2. Audience 15. Palace 7. Cowardly Lion 20. Needs 3. Scarecrow	6.95* 6.62* 3.17 2.69 2.40
12. "The Search for The Wicked Witch". Leave city, Guardian of Gates removes spectacles; in land of West seen by Wicked Witch who sends wolves, crows, etc. to kill; all fail; sends Winkies — slaves — also fail; uses Golden Cap and sends Winged Monkeys who destroy Woodman and Scarecrow; rest prisoners; Witch's greed for Silver Shoes angers Dorothy who tosses water on her and melts her.	6. Animal Attack 14. Silver Shoes 9. Winkies 8. Winged Monkeys 7. Cowardly Lion 11. Emerald City	8.42* 7.57* 4.57 3.00 2.40 2.39
13. "The Rescue". Ask Winkies, freed from Witch, to rescue Woodman and Scarecrow; both saved; Winkies carefully repair Woodman; start for Emerald City; Dorothy takes pretty Golden Cap.	9. Winkies	9.97*
14. "The Winged Monkeys". Lost; call field mice who suggest Golden Cap; call Winged Monkeys and begin flight to Emerald City; on way, Monkeys tell their story and explain charm behind Cap.	8. Winged Monkeys 10. Mice 13. Wizard 18. Pole	7.43* 3.22 2.49 2.29
15. "The Discovery of Oz the Terrible". Back at Emerald City Oz delays seeing them; agrees at threat of Monkeys; discovery of Oz as humbug; he tells of trip via balloon from Omaha; they still believe he can grant their wishes.	20. Needs 13. Wizard 16. Balloon 11. Emerald City	8.38* 4.11* 2.95 2.07
16. "The Magic Art of the Great Humbug". Oz puts pin and needles in Scarecrow to make him "sharp", gives the Woodman a heart inside his chest, feeds the Lion a dose of courage.	20. Needs	2.46
17. "How the Balloon was Launched". Oz and Dorothy try to leave via another balloon they have built. Balloon, with Oz, leaves without Dorothy who is chasing Toto.	16. Balloon	8.10*
18. "Away to the South". They ask Winged Monkeys if they can take Dorothy home; monkeys explain they are powerless outside of Oz; decide to go South to seek aid from Good Witch.	8. Winged Monkeys	3.45
19. "Attacked by the Fighting Trees". Leave city again and start south; in forest are attacked by trees; they escape and continue on.	19. Trees 11. Emerald City	7.92* 2.72
20. "The Dainty China Country". This chapter was not analyzed. Cf. fn. 4.		
21. "The Lion becomes the King of Beasts". Lion is asked by forest animals to destroy monster spider; does so; is elected "king" of beasts.	No factor score value $> +1.99$	
22. "The Country of the Quadlings." Prevented from crossing hill by Hammer Heads who knock them down with their projectile heads; they call Winged Monkeys who carry them over hill to castle of Good Witch of the South.	17. Hammer Heads	9.32*
23. "Glinda grants Dorothy's Wish". 24. "Home Again". (Analyzed together; cf. fn. 4.) Witch arranges for Scarecrow to rule Emerald City, Woodman the Winkies, Lion the animals, returns Golden Cap to Monkeys, shows Dorothy use of Silver Shoes; Dorothy returns home to Kansas.	5. Home	2.01

* Highest score achieved by the factor anywhere in the book.

downcounted 215 list) but on this occasion we approached the list in a completely different fashion. Disregarding *any* attempt at objectivity we carefully selected a set of 215 words corresponding to our knowledge of the book and our hopes as to what factors would be extracted. In short, we loaded the matrix for analysis with the *best* set of 215 words we could possibly choose. From that point on, analysis was conducted according to standard procedures and twenty factors were extracted.

Space does not permit us to present this complete factoring run here. Inspection of the two factor structures, that from the "downcounted" data and that from the "chosen" data was very encouraging; from our knowledge of the book we found no difficulty in matching sixteen of the twenty downcounted factors with those in the chosen word results. Table 7 presents these matchings and indicates those factors that could not be matched from one analysis to the other. While it was clear to us that, for example, downcounted Factor 15 was the "same" factor as chosen Factor 16, it was also clear that this congruence might not be apparent to someone who was not intimately familiar with the book. We therefore searched for a way in which to demonstrate this similarity. Our technique was to take the factor scores for Factor 15 across the entire book and correlate them with the factor scores for the (ostensibly) matched Factor 16. We did this for each of the fifteen unilateral matches in the data and the results were startling. The last column of Table 7 presents these correlations; there is one correlation of .49, one of .79, one more at .86 and the remaining twelve correlations all have values equal to or greater than .97.

Because of the size of these coefficients, we were concerned that some artifact might be operating. We could think of only one: the number of words in common between each two potentially matched factors. Thus, despite our different methods of choice, were a pair of factors yielding high correlations based on basically the same set of highly loaded words (loadings greater than .49), we would be building correlations based mainly on a common set of frequencies.

Table 7 also examines this possibility and demonstrates that it is untenable. Using the downcounted factors as the base, we established the number of downcounted factor-scoring words in common with the possibly matched chosen factor-scoring words. This data is presented both as a fraction and a percentage; inspection of the correlations clearly demonstrates the complete lack of effect of the "commonality" on the correlation.

The only other possibility for an artifact of which we were aware also hangs on the question of common words. Thus, if the total 215 matrix of downcounted data is composed of a high proportion of the 215 chosen word data, its similarity in total words might allow extraction of basically identical factors with the few different words holding onto the high loading areas. It is sufficient to note that only 98 words in the downcounted data and chosen word data are in common, i.e., 46%.

TABLE 7

A Comparison Between Downcounted Word Factors and Chosen Word Factors

Factor mnemonic	Downcount Factor	Chosen factor	Overlap		Correlation
Tin Woodman	1	2	7/14	50%	.99
Audience	2	3	5/11	45%	.97
Scarecrow	3	7,19			
Munchkins	4	5	5/9	56%	.97
Home	5	4	4/10	40%	.98
Animal Attack	6	1	4/8	50%	.86
Cowardly Lion	7	6	4/5	80%	.98
Winged Monkeys	8	20	3/8	38%	.79
Winkies	9	18	2/7	29%	.97
Mice	10	9	3/7	47%	.99
Emerald City	11	12	4/5	80%	.99
Poppies	12	13	4/9	44%	.98
Wizard	13	10	4/8	50%	.69
Silver Shoes	14	14	3/7	43%	.99
Palace	15	16	1/9	11%	.97
Balloon	16	8	1/7	14%	.49
Hammer Heads	17				
Pole	18				
Trees	19				
Needs	20				
King of Beasts		11			
Leaving Oz		15			
Kalidahs		17			

DISCUSSION

The results presented in this paper have led us to four major conclusions. We shall discuss each in turn.

Factor Validity

We raised two questions concerning factor validity in our analysis of the *Oz* data. The first tested validity by asking the extent to which each factor "fit" the data; were the segments for which a factor was most heavily scored good representations of that factor and vice-versa. Secondly, we asked the extent to which the configuration of factors was able to describe the data base itself. Both of these questions were answered affirmatively in the *Oz* data and

while the second has never been tested on any other data-base, the question of "fit" has been examined now over some widely differing sources of material. We take it then that the approach we have pursued is indeed a viable one and that the factors which it extracts are valid representations of the data which elicits them.

Data Base Description

While raised mainly as a validational criterion, the ability of the factors to describe the *Oz* data on a chapter-by-chapter basis has suggested the possibility of a configurational approach towards the description of major content. In such an approach, the information utilized would be the set of factors and their standardized scores operating configurationally over the various units of the data base under analysis; it might thus be feasible to describe changes in a data base by noting the configurational changes taking place. We have done no research into this area as yet but there are several statistical techniques available for configurational analysis and we shall begin to investigate them soon.

Word Selection

The high degree of correlation between the *Oz* downcounted factors as contrasted with the chosen-word factors implies some degree of insensitivity, in the factor *structure*, to the words available for building these factors. We have no information, at present, as to the degree of that insensitivity other than that furnished by the analyses on *Oz*. Thus, we do not know whether the 46% overlap figure between the two sets of words analyzed represents a figure that is adequate because of the nature of the data; it is possible that a much greater overlap may be required when dealing with data whose interrelationships are more subtle and complex than is the case with *Oz*. Further, our UHH rules, operating on the downcount analysis, fairly well restrict the data analyzed to nouns, adjectives, verbs, and adverbs. Whether the kind of concordance obtained with *Oz* would hold when data is analyzed with a heavy preponderance of articles, prepositions, conjunctions, etc., is something we cannot presently answer. Nevertheless, *some* degree of insensitivity is a clearly demonstrable finding and the fact that this robustness does exist has considerable implications for our future lines of research.

We have noted earlier in the paper that the major critical problem we face in the operational procedures of WORDS is in the selection of what words to retain for analysis. We have tacitly assumed that the deletion of words would cause changes in the factors obtained and in the configurational relationships among these factors. Our first indication that this assumed sensitivity might be overstated came when we began our investigation into a UHH rather than an SYN approach. As noted in the paper, both approaches yielded factors many of which could be related one to the other. Several of the factors, however, could not. The *Oz* data confirms this finding. Of the twenty factors extracted in both the downcounted and chosen-word analyses, four of the downcounted factors could not be matched to those in the chosen-word set while three of the chosen-word factors were not matchable; (the discrepancy exists because one of the downcounted factors — the Scarecrow — separated into two factors in the chosen-word data). Both sets of unmatched factors, the three from the chosen-word data and the four from the downcounted data, were equally "good". Table 7 shows that the downcounted data extracted unmatched factors for the Hammer Heads (Factor 17), The Pole (Factor 18), the Fighting Trees (Factor 19), and the Needs (Factor 20). Likewise, the chosen-word data allowed extraction of unmatched factors representing the Kalidahs (Factor 17), Dorothy's flight from the Land of Oz (Factor 15) and King of Beasts (Factor 11). This last factor is worth noting because it will be remembered that no factor in the downcounted analysis obtained a standard score of over +2.00 in Chapter 21, "The Lion Becomes the King of Beasts"; scoring Factor 11 from the chosen-word analysis, however, yields a factor score of +8.00 for Chapter 21. Clearly, then, as we have analyzed the data, a change in the set of words submitted for analysis *does* cause some change in the factors extracted. We are not sure but that some of these "missing" factors might not have been extracted had we continued factor extraction past our limit of 20 but this is problematic. Nevertheless, the fact remains that a considerable change in submitted words still yielded a match on sixteen of the twenty factors, a matched percentage of 80%.

The implication of this result seems clear: We have some margin of safety in our selection of what words will be submitted for analysis. This implication, coupled with the fact that the downcounted data furnished as good a description of the data as did the chosen-word analysis, strongly supports our feeling that a UHH approach, with its advantages of extreme speed and objectivity, is the correct way to pursue our future developments in the system. We shall speak more, later, about some of our plans for increasing the efficiency of the UHH approach.

Inclusion of Non-Word Materials

It will be remembered that we did two analyses on the Gottschalk data presented earlier. In both analyses we submitted the same list of 190 different downcounted words for factor extraction but in the first analysis these words were examined by themselves whereas the second analysis added twenty-five nonverbal variables for analysis; these twenty-five variables were composed of various categories of content analysis, physiological data, etc. The results indicated that the factor structure of both analyses was almost identical insofar as the words and their loadings were concerned. What happened was that the non-verbal data tended to appear within factors already established, in its absence, during the first analysis.

We feel that this finding offers a line of development for the use of WORDS that is quite interesting. We can envision at least two major uses for the inclusion of non-verbal data along with the submitted list of words. On the one hand, we believe it would be interesting to see what descriptive value could be furnished by such non-verbal data in order to add to the utility of the extracted factors and to clarify further the segments of the data base chosen for examination because of high scores on that factor. On the other hand, we can see a use of WORDS in the developmental phases of a categorization system which would allow the developer or investigator of that content-analytic method to investigate the degree to which his content categories are intrinsically related to the various content materials uncovered by the factor structure itself.

PLANS FOR THE FUTURE

Methodologic Issues

There are three major research areas which we intend to pursue and which we should now like briefly to detail.

Statistical Word Selection

The results presented in the paper have clearly suggested that we have some flexibility in the choice of words to be submitted for analysis. We have long been interested in objective methods for such word selection but were troubled because objectivity seemed to demand a price in factor interpretability and meaningfulness. The UHH approach has, however, given us sufficient encouragement to look into this issue further.

In the standard UHH approach as we had formulated it, the final selection of words for analysis was done by downcounting a frequency ordered list of remaining words; this technique, of course, guaranteed that the highest frequency words would be included for analysis with the low frequency words being deleted. There is nothing compelling, however, about such an approach. Rather than depending on frequency selection, we plan to pay extensive attention to that correlation matrix which precedes the factoring run as a method for making word choices. Our reasoning is as follows: An intercorrelation matrix computation is very much faster than a factoring run given the same size of matrix input. With correlations on a 215×215 matrix being computed across, say, a hundred observations, we can reasonably expect the correlation program to run at least eight to ten times faster than the factoring run that will follow it. Further, a factoring run demands more of the machine's available core capacity than does a correlational approach and it is therefore feasible to run larger matrices through a correlation program than through a factor analysis program on a machine of a given size. Putting these facts together, we intend to allow correlational runs on word matrices of orders running to about 800, a size which is usually capable of holding *all* different words left in a data base after deinflection to root form. We shall then utilize another program to inspect this matrix of intercorrelations and to choose from it the 215 words best meeting a set of criteria that will ensure the development of "good" factor structure if, indeed, such factor structure is inherent in the data. There are at least two criteria that make for "good" factors; one, which has been discussed extensively during the paper, is the validity of each of the factors and of the factor configuration. Another, stemming from common factor analytic usage, is simply the loadings on each extracted factor — how much variance do they extract — and, as a result of summation across the factors, how much variance does the total extracted factor set remove from the input matrix. We do not believe that these criteria are independent; we have found, in prior research, that good statistical factors tend to be the more valid factors for our use.

The statistical criterion for factorial "goodness" then is one approach that we can utilize. Since it is a fact that correlational matrices with very high overall correlations will yield better statistical factor structure than those with very low overall correlations, we should like to investigate the possibility that good factor structure can be obtained by eliminating words whose overall correlations are low in favor of those with high mean correlations. Another, non-independent approach, may lie with the variance of the correlations obtained between a given word and all

other words in the matrix. Other things being equal, high variance is better than low for factor analytic operations; such variance is obviously not independent of the mean correlational level associated with a word but may allow selection of words for retention from among other words with equal mean correlations.

Should such an approach prove productive, we would have a completely objective and very fast method for the selection of the "right" words to be included in a factoring run.

Measurement of Specific Words

We are interested in exploring a somewhat different approach towards the "measurement" of specific and key words in a data base. As it is currently, all words are potentially admissable for analysis in a WORDS run. If a word, for example *mother*, is in the data base then it stands a chance of admission for analysis that is independent of its meaning. We believe, however, that if the word *mother* is an important one for the user of the system, it might be fruitful to analyze the data base deliberately leaving out the word in any such analysis. We should then be interested to see what happens to factor scoring techniques as they are applied, for *all* factors, on those observational segments where *mother* does not appear *vs.* those where it does. We have no evidence on what the effect of such an operation will be but think the possibility of success interesting enough to give it some priority in our future research efforts.

Reprogramming of WORDS

The University of Rochester has recently acquired an IBM 360 model 50 computer and will, within another year, update that machine to a model 65. WORDS will be reprogrammed to run on the 360. Programming on the IBM 7074 had, of necessity, to be in assembler language because no higher level language existed capable of doing the job. PL/I has met that need and reprogramming for WORDS will be in that language. Because the 360 is a very popular machine, we shall, for the first time, have the ability and opportunity to make the WORDS System available to others outside the University proper.

While PL/I cannot come even close to matching the efficiency of assembler language coding, it allows us a high degree of programming efficiency and offers the distinct possibility, within the next two years, of being implemented on a number of other manufacturer's machines; this, of course, would allow even further dissemination of the WORDS System. Further, with the increasing speed of third generation machines, the overhead generated by PL/I should be more than compensated for by the increased operating efficiency of the target computer.

In this reprogramming, we shall begin investigation of a data flow logic which we hope to implement. As originally constructed, WORDS was based on the concept of repetitive runs on the computer for the purpose of data reduction with each run taking its input from the prior run's output. With the marked success we have obtained in a UHH approach, with much faster and larger machines available, and with the possibility of word selection being accomplished by a statistical criterion embodied in the correlation matrix, we believe it will be possible to reduce the complete analysis of a data base into two runs on the computer. The first, and more trivial, of these two runs would be for purposes of correcting spelling and any other errors that have crept into the data base during punching and initial entry into the system. The second run would then take place in a fashion somewhat similar to the following: The data would have parts of speech assigned, would go to an analytically oriented deinflection routine, would have words changed and/or deleted according to pre-set rules, e.g., delete all non-verb forms of *like*, and would have all words whose frequency is equal to or less than some pre-set criterion deleted; the remaining words would then be readied for a complete intercorrelation matrix whose results, as mentioned earlier, would be used to select the N highest correlating words for submission to factoring. Results of the factor procedure would be automatically submitted for rotation and the rotated factors would be channeled through for factor scoring on the original data with results of the scoring being made available graphically (for plotting offline) as well as in their usual printed form.

While an automatic procedure of this kind must await the results of our research into the effect of using the correlation matrix as the statistical criterion for selection of the factoring matrix, the programming and systems logic embodied in the preceding description are well within the state of the art of both hardware and software of present third generation machines. Indeed, the entire second run we have described, assuming a thousand different words for initial screening, with final factoring on approximately a 200×200 matrix should run in somewhat less than two hours on the model 65 IBM 360 that will soon be available at the University of Rochester. Thus, the possibility of utilizing WORDS on an almost completely automatic, and therefore almost completely objective, approach towards the analysis of major content cluster is potentially quite feasible.

REFERENCES

Gottschalk, L. A., Ed. *Comparative psycholinguistic analysis of two psychotherapeutic interviews* (New York, International Universities Press, 1961).

Iker, H. P., and N. I. Harway, "Computer analysis of content in psychotherapy", *Psychological Reports*, 14, (1964), 720-722.

——, "Objective content analysis of psychotherapy by computer", in K. Enslein, Ed., *Data acquisition and processing in biology and medicine*, Vol. 4 (New York, Pergamon Press, 1966).

——, "A computer approach towards the analysis of content", *Behavioral Science*, 10 (1965), 173-183.

APPENDIX 1

Current Structure of Words

WORDS currently consists of forty programs. Ignoring those programs which can only be called internally (by another program) and those called by a "package" call (causing internal manufactures of a set calling sequence), there are thirty-two programs available to the user. Each of these programs will be briefly described later. Initially, however, we will detail the basic system-, data-, and program-organization of WORDS.

Systems Organizations

To use WORDS, a series of control cards must be prepared by the user indicating what programs are to be called, in what order, what each program is to do, where each program is to locate and leave its input and output data.

Since preparation of control cards can be complex, it is important that all such cards be extensively screened before allowing a run to begin. This function is satisfied by requiring the first program in any WORDS run to be CHECK. CHECK will subject every control card to an extensive series of generalized validity checks and then further check each card for any idiosyncratic formating peculiar to the particular program being called. CHECK allows the run to proceed only if all cards are error free. After receiving the last control card, CHECK then issues an internal call for the administrative WORDS monitor, MNTRA.

MNTRA uses the control cards, passed by CHECK, and (1) sets up a general calling configuration for the entire run, (2) replaces sort parameter mnemonics with actual sort fields for later use by the sort programs, (3) imposes extensive configurational checks on the run to rule out any logical impossibilities (any of which would then result in a dump of the entire run), (4) forms the entire calling sequence, package generated sub-sequences, sort mnemonics, etc., into a continous block of data which is written to the online 1301 disk file to form a common communications block and then (5) issues a call for the executive WORDS monitor, MNTRB.

MNTRB is normally called on completion of each program in the calling sequence. Upon obtaining control of the machine, MNTRB (1) retrieves the communications block, (2) maintains a record of elapsed time for the just completed program, (3) furnishes the next program in line with necessary data I/O information, (4) indicates where the program may obtain message space if it is needed and (5) where the program's specification card, carrying auxilliary data, may be found; (6) if the program is a sort, transfers necessary sort control information parameters to the core of the machine, (7) updates the communications block and re-writes it back to the disk and (8) issues a call for the next program in the run configuration.

Normally, each program called returns control to MNTRB after completion thus again beginning the series of operations noted. The program PRINT, however, the last called program in every run, does not release to MNTRB but rather to the resident monitor which then terminates the job, tallies total time, and moves on to the next job on the queue.

Throughout the entire run, extensive error trapping procedures are activated. Immediately following CHECK, MNTRA makes certain changes to the core-resident linkages into the resident monitor (later restored by PRINT) in order to prevent normal error recovery under control of the resident. Whenever an irremediable error occurs that is not trapped by a WORDS program itself (typically, I/O or arithmetic), a default branch is taken by the machine operator to keep the queue moving. The changes made to core-linkage by MNTRA overrides this default branch and forces control to another linkage (also provided by MNTRA) which automatically issues a call for PRINT and notes the name of the offending program. On entry, PRINT determines if it has been called normally (by MNTRB) or not. If not, appropriate warning messages of an impending dump are issued and PRINT then goes into normal end-of-job procedures. Inspection of the material furnished by PRINT usually allows the user to diagnose the cause and location of the malfunction. If an error is trapped by a WORDS program, a diagnostic message is issued and the series of events just noted take place by forcing a branch to the PRINT core-linkage routine.

Data Organization

The method for organizing data records in WORDS is dictated by the analytic methods involved in the technique

which requires that the unit of information be the word itself. Thus, each word must constitute an independent and separable machine record.

The standard record format is a collection of eleven fixed length fields comprising thirty characters of information. Each of these eleven fields has a mnemonic associated with it and is referred to by that mnemonic. WORD, a field of fifteen characters, holds the actual english word comprising that record. SPKR, a one digit field, designates the speaker of that word. SPEECH, a one digit code for the part of speech, is inserted by the PARSE series. TIME, a one digit code, is not currently in use. INTV, a three digit number, allows designation of the interview in which the word was found and SEGM, a two digit number, allows referencing the particular observation within that interview. SEQ, a three digit number inserted by the SPLIT program during data input, locates the specific sequential position of the word within the interview/segment combination and SUBSEQ, a one digit number, allows insertion of up to nine words between any two originally input words. SEGTOT, a five digit number indicates total words emitted by the particular speaker in that segment. FREQ, a five digit field, is initially set to one by SPLIT and is then free for whatever use is required by the run. Finally, SPARE, a four digit field, is open to various internal uses by other WORDS programs.

All WORDS programs, other than the initial program SPLIT and the mathematical programs, are written to process records of this length and this format.

Program Organization

WORDS programs belong to one of six functional types: systems control, sorting, editing, record keeping, printing and statistical. Each of these blocks will be described along with a list of contained programs.

Systems Control

These six programs all have the common function of maintaining data flow within the system and between the system and the resident monitor. The block subsumes: CHECK, COPY, FILER, MNTRA, MNTRB, and PRINT.

Sorting

All sort programs are really a third level monitor making use of the same basic applied program, IBM SM148. Calling any sort program actually calls this monitor which brings in the main sort program segments as needed, makes modifications in the segments as required for the particular sort version called, maintains linkages between sort segments and retrieves necessary statistics before returning control to MNTRB. The block subsumes: OMITS, SORTS, and SUMMS.

Editing and Format Manipulation

The nine programs have either the function of changing the data in a file or changing the fields within records in that file or both. Such changes may be accomplished by changing fields, removing or replacing complete records. Programs included are: EDIT, FIXST, PARS1, PARS2, IDIOM, SPLIT, TEXT1, TEXT2, and STRIP.

Record Keeping

The design of WORDS makes it important to maintain a record of changes made to the data. A series of five programs, HSTR1-HSTR4 and HSTRY, process all editing changes before turning control to the actual editing programs. Since I/O scheduling for the series is complex with internally called sort modules, MNTRA accepts a call for HSTRY PKG upon receipt of which it manufactures the appropriate sequence.

Printing

The two modular printing programs, RERYT and PRISM, are used solely to produce output files on the resident monitor's print tape for later listing on an offline 1401.

Statistical

The eight programs in this functional block are designed to carry the reduced WORDS data through an intercorrelation matrix, listing of that matrix, factor analysis, varimax rotation and listing, and finally a scoring procedure utilizing the factor loadings from the varimax data. The programs are: LISTR, CORR1, DUBLR, DECOD, FACTR, VARMX, VDCOD, and SCORE. Since the first four of these programs are constant in any intercorrelational procedure, MNTRA accepts a call for COREL PKG to produce all required calls and I/O.

CHECK. This is the first program called in any WORDS run. Its purpose is to make an extensive series of validity checks on each of the control cards input to the run in order to catch any errors at the beginning rather than the later part of the run.

CORR1. The intercorrelation matrix program of the series. It will handle a matrix of up to 999 variables.

DECOD, DUBLR. DUBLR and DECOD function as paired programs which will almost invariably follow CORR1. DECOD is designed to produce an easily legible output from the CORR1 matrix output by replacing the variable identification numbers by their English words and by allowing an ordering of obtained correlations or by screening them against a pre-set criterion level. DUBLR precedes DECOD and is used to expand the upper symmetric matrix produced by CORR1 into a complete matrix (less the main diagonal).

EDIT. Used to make substantive changes in the data file. It is a very flexible program and allows, among other things, the change of any given word to another, the deletion of any specific word or of all occurrences of that word, the deletion of sets of interviews, speakers, segments, etc. The goal of EDIT is to reduce the total number of different words in the system; in that sense, it is the epitome of the entire system since all reduction changes idiosyncratic to the set of interviews under analysis are accomplished by EDIT.

FACTR. The factor-analytic program of the system. A principal-components algorithm is used to extract up to ninety-nine factors from an intercorrelation matrix of maximum order 215 × 215.

FILER. Allows the production of tape files, all on one tape, designed to serve as future input to the system. Allows the re-input of any of these files on future runs. The program has several safeguards in that files input to the system must be labelled, the input tape is automatically removed after use, the output tape is also removed.

FIXST. Designed for the merging and updating of the striplists used by the STRIP program.

HSTR Series. A series of five programs scheduled and called by use of the HSTRY PKG call. The programs HSTR-1, -2, -3, -4, and HSTRY maintain an accurate record of all data changes made via EDIT. The cumulated history is saved by the FILER program and allows a re-start procedure from an earlier point as well as a history of all changes made to the data.

IDIOM. IDIOM and its second part, PRIDE, will locate idiomatic usages which must be treated differently than regular words since the separate words within an idiom cannot be worked separately. Idiom-constructions, furnished on cards, are located and both listed and punched in a format that is appropriate input to EDIT for any necessary changes.

LISTR. The physical format of records as they are kept by the WORDS system is not appropriate for that set of programs which are designed to mathematically analyze the data. It is the job of the LISTR program to re-format the data when it is finally ready for analysis.

MNTRA, MNTRB. These are, respectively, the administrative and executive monitors of the WORDS system. In brief, MNTRA will accept and process the run control cards which instruct the system as to what programs are needed, when, where, etc., and to construct from this list of cards a calling sequence for the program run configuration; it also makes extensive validity checks on the cards. When MNTRA has constructed this calling sequence it turns control to MNTRB which then takes over the actual calling of each program in the sequence and the task of maintaining adequate communications between programs.

OMITS. A sorting program with the facility for deletion of records which are equal to each other, according to furnished parameters upon which equality is to be assessed, leaving only the first of such records intact. Thus, were the user to want a list of every different word in the data, use of OMITS on the sorting parameter of the word itself will cause deletion of all words which are alike save for the first in the string; the remainder then becomes one record for every different word in the system.

PARS1, PARS2. These programs will insert a part of speech code into each record (word) in the data. PARS1 operates mainly on a dictionary lookup basis, although some logical manipulation is done, in order to make assignments where the grammar code is unambiguous. After resorting the output from PARS1, PARS2 takes over in order to assign the remaining codes according to fairly extensive analytic rules. The assignment of parts of speech is important in reduction since it allows reduction by rule (e.g., "delete all articles and conjunctions") via the EDIT program and in that it allows combinations of words with other words because the specific meaning of the word is defined with its part of speech, e.g., *like* = *enjoy* vs. *like* = *similar.*

PRINT. A dual purpose program. PRINT is called in order to terminate the WORDS system control over the computer before returning the machine to the resident monitor. Before doing so, however, PRINT will compile a record of statistics and messages produced during the actual computer run itself.

PRISM. This is the main printing program of WORDS. PRISM is designed to allow the printing of those files which have been selected for output in order to provide

either a record of certain of the results of that run or to provide an indication of information for planning future runs.

RERYT. Where PRISM is primarily designed to produce lists of any data file contents, RERYT is specifically intended to produce a printed copy of the interviews under analysis in a format similar to that of the orginal typescript. Thus, each speaker is separated from every other, periods are restored to the end of sentences, spacing separates segments, etc. RERYT is used when it is desirable for the user to inspect the "state" of the data after various transformations have been made. After an extensive EDIT run, for example, it is useful to be able to read the interviews in their present form to determine "clinically" just how much meaning is still being retained in the data.

SCORE. Using the correlation matrix input data prepared by LISTR this program accepts a deck of cards — one for each word in the matrix — punched with the varimax loadings for each word on each of the factors on which it loads highly. By using the frequencies of occurrence and the loadings as multipliers, it computes a factor score for each factor in each observation. It then produces a printing file with the factor score means and standard deviations across the entire data set and then lists, for each observation, the raw score and standard score of each of the factors.

SORTS. A sorting program which, unlike SUMMS or OMITS, makes no physical change to the total file in terms of deletion or summarization. SORTS will simply reorder the records in the file into whatever order is specified by the sorting parameters; other than re-ordering, no changes are made.

SPLIT. This program serves as the entry point for raw data into the system. When an interview is originally punched, as many words are placed on each card as is feasible. These cards are then put on tape either via an offline 1401 operation or online by COPY. SPLIT takes this card image tape as input and produces a separate WORDS record for every word on every card of the input data. SPLIT will also insert within each record all necessary data for determining the origin of the word, i.e., segment number, interview number, speaker, etc. In addition, it will also assign a sequencing number to each word as a function of its position in the segment in which it was found. Use of these identifying origin data allows SORTS to restore the data to its original order no matter how it has been re-ordered by any other program.

STRIP. Like EDIT, STRIP is designed to make substantive changes to the interviews under analysis. Where EDIT makes such changes on the basis of the specific interviews being handled, STRIP is designed to be applied to every set of interviews that comes along. STRIP may be considered as a de-inflection program whose task it is to place the words in the data into their root form. STRIP, unlike EDIT, has no flexibility in terms of options. It can only replace a given word with another as this replacement is specified by a deck of pre-punched cards. No deletions or other types of changes than replacement, are permitted.

SUMMS. Like OMITS, SUMMS deletes records equal to each other so that only one record of each type (cf. OMITS) remains. Unlike OMITS, however, SUMMS first adds the frequency data of each record deleted to the frequency data of the first record in the string. Thus applying both OMITS and SUMMS to the same data with the same sorting parameters would yield the same set of records on output but SUMMS would have accumulated within each record the summed frequency of all the records deleted.

TEXT1, TEXT2. A pair of programs designed to provide a KWIC type of listing with the programs producing a record for each input word which reports the two words preceding and following that specific word and reports also the interview, segment, and sequence numbers as well as the part of speech for the key-word. Unlike IDIOM which searches for specific idiomatic constructions and then reports both the idiom and the sentence which it contains, the TEXT-programs are designed more as a dictionary producer which allow the user a "random-access" approach so that *any word* can be located in context at any time.

VARMX. The varimax rotation program to be applied to output from FACTR. The program will rotate any set of up to thirty-three factors from the FACTR output tape to a criterion of simple structure. The set of factors to be rotated are selected from the input set by control card punching.

VDCOD. This program serves VARMX as does the DECOD program CORR1. It allows production of an easily legible listing from VARMX with each variable number being replaced by the corresponding English word under analysis, with all factor-loadings ordered by absolute value and with a listing of communalities and variance proportions attributable to each factor.

APPENDIX 2

Factors Extracted from Wizard of Oz *Using Frequency Selection Choices*

Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
ax	91	*Oz (lady)*	94	*farmer*	86	*Munchkins*	82	*Uncle Henry*	92
oil	91	*Oz (head)*	90	*brick*	84	*Witch*	82	*house*	90
Tin Woodman	86	*kill*	85	*Scarecrow*	82	*East*	80	*Aunt Em*	89
tin	86	*send*	83	*road*	66	*woman*	75	*bed*	74
leg	85	*lovely*	78	*stuff*	66	*old*	73	*small*	74
right	79	*eye*	67	*feel*	62	*little*	71	*sun*	72
arm	77	*do*	63	*eat*	60	*wear*	61	*door*	71
body	72	*help*	62	*yellow*	59	*live*	51	*laugh*	67
soon	70	*answer*	58	*walk*	58	*set*	51	*look*	52
girl	65	*throne room*	53	*hurt*	57	*face*	49	*reach*	51
head	60	*surprise*	52	*straw*	56	*Dorothy*	48	*Toto*	41
once	55	*will*	45	*place*	52	*people*	48	*run*	41
work	54	*no*	41	*few*	50	*Good Witch*	44	*middle*	39
tinsmith	50	*tell*	39	*no*	50	*can*	−41	*sit*	38
can	42	*Oz*	38	*brain*	49	*grow*	39	*back*	−33
grow	37	*West*	38	*other*	47	*land*	39	*first*	33
far	−36	*many*	36	*man*	39	*Cowlardly Lion*	−38	*land*	33
help	33	*straw*	34	*Toto*	38	*dress*	37	*grass*	32
old	33	*die*	31	*crow*	37	*country*	35	*eye*	31
come	32	*great*	31	*do*	37	*great*	−34	*hand*	31
one	32	*grow*	31	*leave*	36	*house*	32	*one*	31
put	32	*return*	31	*mind*	35	*look*	32	*ask*	−30
return	31			*right*	33	*the group*	−31	*fall*	30
face	−30			*know*	32			*hard*	30
look	−30			*number*	−32				
				soon	32				
				keep	−30				
				pole	30				
% Variance	5.70		4.80		5.20		4.70		4.80

Factor 6		Factor 7		Factor 8		Factor 9		Factor 10	
wolf	90	*coward*	88	*Gaylette*	93	*Winkies*	67	*mouse*	91
lie	78	*near*	75	*Quelala*	84	*tinsmith*	65	*Queen Mouse*	90
crow (noun)	77	*Cowardly Lion*	67	*time*	82	*set*	54	*safe*	75
die	77	*heart*	62	*Winged Monkeys*	81	*careful*	52	*turn*	60
lay	65	*know*	54	*Golden Cap*	74	*night*	51	*all*	52
number	65	*Toto*	41	*call*	59	*day*	50	*grass*	51
come	51	*stuff*	41	*fly*	58	*tear*	50	*try*	50
tear	51	*try*	39	*next*	51	*pretty*	48	*field*	49
Wicked Witch	48	*great*	37	*wish*	42	*leave*	−43	*run*	44
fly	48	*fast*	36	*the group*	40	*like*	−43	*hurt*	41
Winkies	46	*big*	35	*field*	40	*forest*	−42	*fast*	39
straw	43	*no*	33	*lose*	37	*Good Witch*	−40	*come*	38
Golden Cap	42	*return*	32	*together*	−35	*basket*	39	*near*	37
one	42	*run*	32	*glad*	33	*rule*	−39	*work*	36
foot	39	*beast*	31	*land*	32	*work*	38	*bring*	35
stand	36	*tell*	−31	*good*	31	*keep*	37	*open*	33
next	32	*tin*	31	*sure*	31	*last*	37	*far*	32
ask	−30	*reply*	30	*once*	30	*mind*	37	*live*	31
time	30					*yellow*	36	*speak*	31
						few	35	*big*	30
						start	35	*yellow*	30
						bring	34		
						hand	34		
						stand	−33		
						friend	32		
						live	−32		
						beast	−31		
						ask	−30		
						lay	30		
% Variance	4.30		3.50		4.20		3.70		3.70

Factors Extracted from Wizard of Oz *Using Frequency Selection Choices*

Factor 11		Factor 12		Factor 13		Factor 14		Factor 15	
spectacles	87	*flower*	90	*terrible*	63	*Silver Shoes*	83	*room*	86
Guardian of the Gates	86	*stork*	88	*man*	58	*end*	76	*green*	85
want	75	*sleep*	82	*home*	−56	*water*	70	*soldier*	70
Emerald City	70	*bright*	56	*promise*	54	*Wicked Witch*	65	*dress*	67
bright	56	*fast*	56	*please*	53	*foot*	59	*morning*	65
long (adj.)	−45	*last*	54	*think*	53	*power*	57	*wait*	65
glad	41	*carry*	52	*stand*	52	*use*	51	*girl*	59
first	40	*find*	51	*voice*	51	*take*	46	*see*	51
wish	40	*like*	50	*little*	41	*Dorothy*	44	*bed*	50
sun	39	*fall*	45	*Oz*	40	*hard*	41	*throne room*	48
surprise	39	*let*	44	*speak*	37	*begin*	37	*night*	48
man	34	*hand*	42	*one*	36	*bring*	−35	*door*	47
speak	34	*wait*	40	*back*	−35	*beauty*	−30	*pretty*	43
eat	33	*must*	79	*rule*	−33	*open*	−30	*pass*	39
open	32	*take*	−38	*forest*	−32			*big*	38
may	31	*river*	35	*must*	32			*silk*	38
night	31	*field*	34	*Kansas*	−31			*get*	−37
beast	−30	*few*	31	*beauty*	−31			*course*	36
give	30	*lovely*	31	*first*	−31			*middle*	36
great	−30			*friend*	31			*one*	36
like	30			*wait*	30			*can*	−34
								speak	34
								begin	−33
								cry	30
								eye	30
								many	30
								wear	30
% Variance	3.70		4.10		3.10		3.30		4.40

Factor 16		Factor 17		Factor 18		Factor 19		Factor 20	
balloon	84	*pretty*	55	*pole*	76	*tree*	82	*courage*	74
air	71	*Hammer Heads*	54	*river*	75	*side*	63	*real*	64
silk	71	*kind*	53	*middle*	60	*branch*	57	*brain*	61
make	64	*dress*	43	*let*	48	*seem*	53	*very*	61
basket	63	*reach*	41	*get*	45	*other*	49	*many*	57
lose	55	*rest*	41	*water*	44	*walk*	45	*use*	53
go	52	*will*	41	*land*	40	*journey*	44	*sure*	50
day	45	*indeed*	−40	*rest*	40	*first*	43	*give*	49
get	45	*back*	39	*fast*	39	*real*	−43	*reply*	49
now	45	*grow*	−39	*West*	34	*the group*	42	*day*	47
should	45	*thank*	39	*leave*	34	*must*	42	*find*	47
together	45	*take*	37	*animal*	−31	*long* (adj.)	41	*heart*	46
tear	38	*field*	36	*great*	−31	*thank*	−41	*think*	45
people	37	*head*	36	*begin*	30	*turn*	41	*fear*	44
find	34	*country*	35	*may*	30	*forest*	39	*people*	44
will	34	*friend*	35			*Kansas*	−38	*Oz*	40
Kansas	33	*pass*	34			*kind*	−38	*put*	37
thank	32	*course*	−33			*next*	37	*big*	−36
Oz	31	*sit*	32			*tell*	−37	*good*	34
last	31	*voice*	31			*rest*	36	*face*	33
		must	30			*bring*	−35	*live*	32
		open	−30			*course*	−35	*may*	30
						answer	−34	*morning*	30
						surprise	34		
						Good Witch	−31		
						beast	31		
						look	30		
% Variance	3.70		2.90		2.90		4.10		3.80

Review Articles

Gerald Lefkoff (ed.), *Papers from the West Virginia University Conference on Computer Applications in Music.* Morgantown, West Virginia University Library, 1967. 105 pp.

This slender volume reports the substance of a conference at West Virginia University on April 29-30, 1966. The papers included are by Barry S. Brook, "Music Bibliography and the Computer", Allen Forte, "Computer-Implemented Analysis of Musical Structure", Gerald Lefkoff, "Computers and the Study of Musical Style", Lejaren A. Hiller, "Programming a Computer for Musical Composition", and Charles C. Cook, "An Introduction to the Information Processing Capabilities of the Computer". At first blush, it would seem that papers given at a 1966 conference are ancient history, but not so in this instance.

The contribution by Cook, a general but pithy introduction to computers, serves to initiate the novice and might have better stood as the first article in the book. It is a clear and warm invitation to humanists, and to musicians in particular, to use the machine but it is also a clear warning that old modes of thought and fuzzy thinking will have to be discarded. Allen Forte's paper also invites the scholar and might serve as a model *caveat*. Without belaboring the point, Forte emphasizes the broad capabilities of the computer and its potential strengths in storing and manipulating large quantities of data. In describing one of his own projects, analysis of certain atonal music, the author gives the novice a good deal of insight as to how an expert goes about encoding data and writing a program. Forte employs for encoding the musical score a system developed by Stefan Bauer-Mengelberg which calls for total representation of the score in a non-interpretative manner, an obvious advantage since "... the responsibilities of the encoder are minimal".

Barry S. Brook has also made a contribution to the problem of encoding music through his "Simplified Plaine and Easie Code System for Musicke", and he comments on it at length in this volume. Though possessing many desirable characteristics, especially from a library viewpoint, it seems to me that the Bauer-Mengelberg system has greater capacity to accurately reflect the full musical score. Both of these systems and the abbreviated system discussed by Lefkoff are, I remind music scholars, almost entirely restricted to music since about 1500. Perhaps the difficulties of earlier music notation, and of *current* music notation, for that matter, are so great that we shall never have an all-inclusive encoding procedure.

The first part of Brook's paper is, I am happy to report, completely out-of-date: he has succeeded in his attempt to bring some order to music bibliography through establishing, on an international, cooperative basis, *Répertoire International de la Littérature Musicale* (*RILM*), an abstracting service for current musical scholarship. Brook describes the hopes and procedures of *RILM* and, to a large degree, they are currently being realized. (At a recent joint congress of the International Association of Music Libraries and the International Music Council several fruitful meetings were held to explore the problems *RILM* faces and how they can best be met.)

Lejaren Hiller has been in the forefront of musicians using the computer for musical composition and, in my opinion, a refreshingly conservative musical aesthetic still shows in this experimentally minded composer. His discussion of their experiments at the University of Illinois, though perhaps too highly detailed, is clear expository writing, a virtue in these troubled times. He dwells at several points on the stochastic process, a process treated in detail and at great length by a contemporary composer, Iannis Xenakis. (Xenakis has published his own thoughts and mathematics in *Musiques Formelles*, a special issue of the journal, *La Revue Musicale*, Paris, Richard-Masse, 1963.)

In sum, then, the papers are a helpful addition to the growing literature dealing with computers in the service of musicians.

University of North Carolina JAMES W. PRUETT

Burton R. Pollin, *Godwin Criticism: A Synoptic Bibliography*. Toronto, University of Toronto Press, 1967.

Burton Pollin's bibliography of Godwin is a fascinating book. It is a tribute to the tireless industry of the compiler, who is a Professor of English in the City University of New York and who has himself contributed a dozen studies of Godwin, including his dissertation, *Education and Enlightenment in the Works of William Godwin*, which earned him a doctorate from Columbia University. *Godwin Criticism* is also a demonstration of the advantages of using a computer in such work. As far as I know, it is the first computerized bibliography of an author. Thus it deserves our attention, even if we are lukewarm about Godwin. Personally, I find that Godwin comes rather low on my list of active interests, but I am surprised to discover by a perusal of this book to what extent he seems to have been on the minds and in the books of his contemporaries and their descendants. A true bibliography (a list of works by or about an author) would hardly reveal this so clearly. Only this kind of book, which is both a bibliography of works about Godwin and an allusion-book,[1] is able to show it.

The main part of this book is a listing of 3379 items — the listing runs from 1-4214 with gaps for additions — each one consisting of a bibliographical description and an informative note. There are two sets of parallel lists: periodicals and books dealing with the period 1783-1836 (Godwin's first publication to his death); and periodicals and books from 1837 to 1966. Within each of the four sub-lists, the ordering is alphabetical by title of periodical and author, respectively. Each of the entry numbers is followed by a letter identifying the item as a book or an article about Godwin, a review of one of his works, a review of a book about him, an obituary, or a passage about him in another work, a mere comment, or a quotation from Godwin or a mention of his name.

The notes, which attest to the fact that Mr. Pollin personally examined nearly all of this huge mass of material, consist primarily of a description of the work and a summary of the contents. As much as possible he quotes or paraphrases the words of the original when this is likely to be of interest, as in the early and contemporary comments on Godwin. The notes are informative and, as far as can be told, objectively fair in their appraisals, as evidenced by the inclusion of unfavorable criticism in two reviews of Pollin's own book (2272, 2446). The failure clearly to distinguish between quoted material and the annotator's own words derives from the limitations of the 026 keypunch, which is not equipped with single and double quotation marks.

It is obvious that at the present stage of availability of machine-readable material in libraries, the computer could have been of no help to Mr. Pollin in turning up any of this material. Even as indefatigable an investigator as he is needed three years to compile this work. The list of libraries to which he acknowledges indebtedness takes half a page (p. xxxviii) and reveals the extensive peregrinations imposed on the modern bibliographer (Boston, Brussels, London, Oxford, Cambridge, Madrid, Chicago, Washington, Philadelphia, Paris and New York). The research, in other words, was done in the traditional way: search, collect, transcribe and organize. These aspects of the bibliography reflect Mr. Pollin's ability and his conception of what a reference work should be. On this plane, the enterprise deserves great credit, for the kind of information that has been placed at our disposal here could not be duplicated elsewhere. Whether the scholar prizes Godwin or not, he cannot help feeling satisfied that such a compilation exists.

The contribution of modern data-processing to the making of this book is another matter.[2] As Mr. Pollin explains in his full Introduction (p. xviii), some special aspects of the form of his work are to be attributed to the exigencies of the computer and the convenience of the key-punch operators. The most visible symptom of computer processing is the uniformity of the type: everything is upper-case; there are no italics or bold-face to provide typographical emphasis and increased readability. The bibliographical entries and some of the notes are broken into fields by virgules (slashes) to make possible manipulation of parts of the entries, particularly for indexing. Such a system makes the bibliographical style diverge somewhat from conventional norms of bibliographical description, especially for periodicals. Mr. Pollin does not state what kind of factors, whether technological or bibliographical, underlay his decision to arrange all his periodical items alphabetically under the name of the periodical. Such a system can be justified, since it offers a way of handling the anonymous authorship of early periodicals without recourse to "Anon.", and since the author's names can be retrieved by reference to the appropriate Index. It does, however, give us five consecutive pages from *Notes & Queries* and three and a half from *The Times Literary Supplement*. Since the basic order is chronological, it might have been best to

[1] The classic example of this type is Caroline Spurgeon's *Five Hundred Years of Chaucer Criticism and Allusion: 1357-1900* (Cambridge, 1925).

[2] As a symbol of its autonomy, the computer has inserted an unsightly system message in job control language on p. 557, appropriately enough at the end of Index I ("Persons Mentioned").

arrange both the periodical items and the books chronologically. After all, the other information can all be dug out of the Indexes.

It is in these Indexes that the technological advantages of modern data-processing are most evident and it is they which most obviously constitute the contribution of the machine. At the cost of tailoring his input format to the needs of the machine, Mr. Pollin has rewarded the user with a wealth of indexes. To the scholar who is resigned to using books (even bibliographies) with one incomplete and unimaginative index, Mr. Pollin's eleven indexes represent the fulfillment of a dream. There is an index of persons mentioned in the notes and a similar one of books, both selective, to be sure, but very useful. There is, of course, a complete index of the authors of the works listed in the bibliography. Index IV is a chronological listing of all the items, providing the information for charting scholarly interest in Godwin at various periods. Index V lists the entry numbers for the reviews of Godwin's works under their titles, arranged alphabetically. Index VI is a selection of the more important items on Godwin, arranged under three categories (books, articles, necrologies) and listed by entry number. The first category reprints the whole bibliographical entry, but the second and third are merely blocks of entry numbers and thus of limited usefulness. Index VII gives the entry numbers of all the reviews of books *about* Godwin under the names of the authors. In Index VIII, we are provided with a list of all the entries in languages other than English (Dutch to Swedish, with French predominating). Totals would have been useful here. Index IX provides an alphabetical list of the periodicals annotated in the two periodical lists, with the span of entry numbers covered by each. There we find that the *Poughkeepsie Casket* and the *Prairie Schooner* were each drawn on for a necrology and short comment, respectively, whereas the *Critical Review* furnished forty-seven items and the various publications of Keio University (Japan) fifteen items. Index X is devoted to statistics: (1) frequency distribution of items according to descriptor (e.g., twenty-two necrologies, fifty books or pamphlets solely about Godwin); (2) numerical contents of each index; and (3) a graph of publications about Godwin year by year, and by decade, with each (or ten items) represented by an 'X'. In recent times, 1951 and 1953 seem to have been vintage years for Godwin-fanciers, doubtless as a result of Mr. Pollin's work in this field. The book concludes with a list of Godwin's own works in chronological order with reprints and translations noted.

Dr. George W. Logemann has contributed a compact technical introduction ("Programming the Book"), concisely setting forth the machine aspects of the book and clearly showing the contribution of the computer to each aspect of the work. These eight pages should serve as a valuable guide to future bibliographers interested in making their productions better reference works, more helpful to the user. For students of Godwin and the Shelley circle, *Godwin Criticism* will henceforth be a necessity. For the rest of the scholarly community, it can stand as a model of the valuable collaboration of man and machine.

Teachers College, Columbia University LOUIS T. MILIC

Gordon R. Wood, *Sub-Regional Speech Variations in Vocabulary, Grammar, and Pronunciation* (= *Cooperative Research Project*, No. 3046). Edwardsville, Illinois, Southern Illinois University, 1967

There can be no doubt about the need for electronic data processing in a discipline as prone to collecting as dialectology is today. Dr. Wood's work, for this reason alone, should be considered carefully by dialectologists as they contemplate future projects.

Wood's corpus consists of 33 tape-recorderd inteviews of white, native born, 20- to 60-year-old informants from Tennessee, Mississippi, Georgia and Alabama. A modified version of Sapon's *A Pictorial Linguistic Interview Manual* was used to elicit running speech. Each recording was typescripted into uncontracted and conventionally punctuated English sentences. For this body of data, three programs (lexical, phonological and syntactical) were written and run.

The lexical study showed a dissemination of known Midland vocabulary into the Southern region, somewhat less influence of Southern words on the Midland area and certain obsolescence and innovation of a commercial sort. Wood's handling of lexical data seems adequate and enlightening.

The phonological and syntactical programs, however, may be criticized on several scores. After making a narrow transcription of specific words in the lexical program along with certain others in the running speech, Wood replaced this transcription with a broad one "... nearer to phonemic in the linguistic spectrum" for simplicity of coding. The coding system, for example, allows neither for fronted, backed, raised, lowered, nasalized or lengthened vowels nor for fronted, backed, lengthened, aspirated or devoiced consonants. There is no provision for glottal stops, stress, pitch or such essential distinctions as those between diphthongs and on or off glides. Such editing

must surely have overlooked some important characteristics of Southern speech, particularly between social classes or sexes. By simplifying his coding operation in this way Wood may well have limited the inventory of possible contrasts which dialectologists pride themselves on finding. But despite this self-imposed limitation, he is able to generalize about the phonology of much of the area under consideration.

To study syntax, Wood first searched the typescripted text for previously tabulated lexical items. That word was then considered some sort of sentence base, and the preceding and following words were assigned slots by contiguity to the base. From these tabulations, Wood studied the positional frequencies of selected parts of speech as they occurred in the various preceding and following slots. He then tabulated positional frequency of occurrence of these parts of speech by counties. Not surprisingly, Wood found little evidence to support any claim that there are geographical sub-areas within his area which have significant syntax differences. To suggest that it might have been more useful to have studied syntax differently would be to beg the question. It would appear to be more useful to examine multiple embedding, coordination versus subordination or various kinds of transformations.

In terms of his approach to the problem of information retrieval, then, Wood is by far more successful in his lexical program than in phonology or syntax. Other shortcomings are not attributable to Wood alone. The linguistic atlas social categories have been under fire from sociologists for some time and, of course, convenience sampling of informants is a highly suspect procedure. But Wood has made a notable contribution to the complex problem of data handling of dialect information and his work will serve as a useful empirical model for further research and improvements.

Sociolinguistics Program ROGER W. SHUY
Center for Applied Linguistics
Washington D. C.

Contents

Some articles for forthcoming issues:

EVERETT ALLDREDGE, Preservation of Documentation for Conventional and Automated Systems

GARY BERLIND and GEORGE W. LOGEMANN, An Algorithm for Musical Transposition

ERWIN DANZIGER, Tutorial on Computer Hardware

ALLEN FORTE, Analytic Methods of Machine Storage

DOUGLAS HINTZMAN, Learning and Memory in a Discrimination Net

OLE R. HOLSTI, Computer Content Analysis for Measuring Attitudes: The Assessment of Qualities and Performance

CHARLES H. KELLOGG, Data Management in Ordinary English: Examples

ALASTAIR MCKINNON and ROGER WEBSTER, A Method of 'Author' Identification

R. NARASIMHAN, Intelligence and Artificial Intelligence

ELLIS B. PAGE, A Program for the Automatic Evaluation of Student Essays

JAMES B. RHOADS, The Role of the National Archives in Facilitating Statistical Inquiry

WALTER A. SEDELOW, JR., History as Language (Part I); Verbal Structure of Hume's *History of England* (Part II)

HUNG-CHING TAO, A Chinese Computer Alphabet for Automatic Machine Processing

WAYNE TOSH, Machine Translation, 1969

Dictionnaire inverse de la langue française

par

ALPHONSE JUILLAND

(*Janua Linguarum, Series Practica*, 7). 1965. lx + 504 pp., 9 figs. 20 × 27 cm. Cloth Dglds. 90,—/$25.75/125F

The Dictionary contains more than 40,000 words listed in inverse alphabetical order. All words contained in the *French Oxford Concise Dictionary* appear in the *DILF*, and also some which do not.

The inverse ordering of the words is based on phonetic, not orthographic transcriptions, so that French words ending in the same sound or sounds are grouped together, even if their spelling is different. This also means that, up to a point, the *DILF* can also serve as a dictionary of French rhymes.

The dictionary consists of an "Introduction" (pp. i-xvi), an extensive "Table des Matières" (pp. xvii-lx), the "Dictionnaire" (pp. 1-434), and three "Appendices" (pp. 435-503).

The introduction briefly discusses the general character of inverse dictionaries and justifies the particular features which characterize the *DILF*: the inventory of phonemes used in the transcription of French words; the order of phonemes underlying their inverse ordering; the outstanding problems encountered in the phonemic interpretation of French words; and the principles that have governed the selection of the words introduced in the dictionary.

The extensive table of contents facilitates the use of the principles that have governed the selection, giving the usual references to the page where words ending in a given phone, diphone, or triphone, are to be found. It also provides the number of words in the *DILF* which end in a particular sound or sequence of sounds.

The body of the *DILF* consists of thirty-two chapters, each of which lists in inverse alphabetical order the words ending in one of the thirty-two sounds found in the final position of French words (the vowels /ɔ/ and /œ/ and the semi-vowels /w/ and /ɥ/, do not occur word finally). Each chapter is sub-divided into sections relative to the penultimate sound of French words: e.g., words ending in *-a* are sub-divided into words ending in *-aa*, *-ba*, *-ka*, *-da*, etc., the diphone that introduces each section being printed in bold characters in the middle of the column. Each section is further sub-divided into paragraphs relative to the ante-penultimate sound of each word, e.g., words ending in *-ba* are further sub-divided into words ending in *-aba*, *-eba*, *-iba*, etc. Paragraphs are separated from one another by a blank line and the appropriate triphone is marked in the column on the right of the phonetic transcription of the first word in each paragraph.

Those interested in words ending in a certain phone, e.g., *-a*, can simply check the chapter headings which always begin on a new page; those interested in words ending in a certain diphone, e.g., *-aba*, can check the paragraph headings printed in normal characters after the phonetic transcriptions of the first word in the paragraph.

Every entry in the dictionary consists of information distributed in the two columns: in the left column, the orthographic transcription followed by its part-of-speech abbreviation (*v.* for verbs, *adj.* for adjectives, *m.* for maculine nouns, *f.* for feminine nouns, etc.); on the same line but in the right column, the phonetic transcription.

The three appendices summarise the results in the form of statistical hierarchies of the endings which characterise the words in the dictionary. The first appendix is devoted to a study of the last sound of French words, which serve as chapter headings in the dictionary; the second appendix is devoted to a study of the last two sounds, which serve as section headings in the dictionary; the third appendix is devoted to the study of the last three sounds, which serve as paragraph headings in the dictionary. All headings consisting of one sound, two sounds, and three sounds are listed in decreasing order of dictionary frequency (the number of words in the dictionary that end in that particular phone, diphone, or triphone). Each ending is followed by its absolute frequency (number of words), by the relative frequency (in percentages), and by the cumulative frequency at its rank in the hierarchy (in percentages). In order to provide a comparative insight into the structure of French word-finals, the appendices also provide the equivalent figures for the Italian language, and the two sets of data are plotted against one another in appropriate graphic representations.

MOUTON · PUBLISHERS · THE HAGUE / THE NETHERLANDS