VOLUME I | NR. 3 | OCTOBER 1968

COMPUTER STUDIES

in the Humanities and Verbal Behavior

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.

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

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	*INTRAPRSNL SCHIZ WITHDRWL	−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".	5. Home	2.01
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.		

* 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 $\times$ 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 re-order 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