A COMPUTER APPROACH TO CONTENT ANALYSIS: STUDIES USING THE GENERAL INQUIRER SYSTEM

Philip J. Stone, Harvard University
Earl B. Hunt, University of Sydney

The General Inquirer[1] is an IBM 7090 program system that was developed at Harvard in the spring of 1961 for content analysis research problems in the behavioral sciences. The first part of this paper describes this system and how it has been used. During the summer of 1962, the General Inquirer was merged with the Hunt Concept Learner[2, 3] to produce a method for automatic theme analysis[4]. The second part of this paper discusses the rationale behind this development and some recent signs of its future promise.

Within the behavioral sciences, much of the raw data to be analyzed consists of written text. A psychologist, for example, may hand you an inkblot and ask you to describe what you see. A public opinion interviewer may ask for your free answer to his questions. A sociologist may record conference group processes and make transcripts of tape recorded sessions. A political scientist may collect diplomatic notes. In each case, the data is the same: written material such as you are reading right now. From the viewpoint of the behavioral scientist, this is raw data, consisting of words and punctuation marks recorded on a page. The analysis is yet to be done.

As defined by Berelson[5] in 1952, "Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication". Berelson uses the term "objective" to indicate that the procedure should be explicit, one that can be replicated exactly by other analysts. "Systematic" means that "*all* the relevant content is to be analyzed in terms of *all* the relevant categories" in order to secure unbiased information for the hypotheses being tested. "Quantitative," of course, refers to the process of counting positive or negative instances. Finally, in order to do all this, the procedure has to be based on the "manifest" aspects of the text; however, as Berelson points out, "the results of content analysis frequently serve as a basis for the 'interpretation' of latent content".

The General Inquirer was developed to help further the rigor of these procedures. In describing a content analysis procedure to a computer, nothing may be left implicit for processes of intuition. The program is an "objective" description of the content analysis process, it must operate on "manifest" features of the text, and it can be designed to yield carefully counted "quantitative" results. Once the computer handles the task, the program is "systematic" in carrying out all of the details of the analysis. Since a systematic program may analyze many variables at once, one may discover trends that are indeed "latent" to casual observation.

Verbal text may be regarded as the result of many different psychological processes acting at once, each revealing its influence in the final product in different subtle ways. Our classification procedures in the General Inquirer are concerned with recurrently expressed or assumed values, underlying types and intensities of motivation, perceived demands of the environment, and institutionalized structuring of both

"This work was supported by NIMH (JSPH M-4169) and ONR (TOSK 047-003). IBM 7090 computing has been done with the cooperation of the computation centers at MIT, Harvard, U.C.L.A. and Stanford."

demands and action as these may be shown directly or indirectly by text materials.

One common purpose of the General Inquirer is to find psychological indexes that will discriminate between two text sources. While making such discriminations is indeed a problem in identifying authorship, it differs from the classical authorship problems (Federalist papers, Shakespeare-Marlowe, books of Homer, etc.) in several respects. First, each source is usually not a single author, but rather a group of people. Second, our goal is not just to make an efficient discrimination, but is also to gain further understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written. If our only goal was to identify authorship, usually we could find our best clues in idiographic stylistic differences (noun declension preferences, "while" versus "whilst,"[6] etc.).

PROCEDURES

Text preparation. The alphanumeric characters of the text (including punctuation) are keypunched on IBM cards, each card roughly corresponding to a typewritten line with 80 space fixed margins. Between-card breaks come either between words or after a hyphen as in regular typewriting.

System operation. The text is transferred from the punched IBM cards to magnetic tape by an IBM 1401 machine. The 7090 computer then reads the alphanumeric characters and separates them into words and sentences. Certain regular word endings are removed and each word is looked up in a dictionary. If a word is found in the dictionary, tags indicating the word's membership in one or more categories specified by the investigator are attached to the sentence. If a word is not found in the dictionary, it is put on a leftover list for further examination by the investigator. The sentences, together with their tags, are then stored on binary tape for repeated use in inquiry procedures. The investigator uses inquiry procedures to ask for the number of times (and possibly the retrieval of each instance) where a particular combination of tags or specific text words co-occurred in the same sentence. Up to a hundred such questions may be processed on one pass of the binary tape. By arranging the questions into "question sets", it is possible to ask about various disjunctive relationships.

If the keypunched text is marked with a simplified form of syntactic coding, the investigator can make inquiries not only about the cooccurrence of certain text words and/or tags, but also can specify the syntactic relationships that must appear between them. The computer can also use the syntactic codes to help tag correctly many otherwise ambiguous words.

A General Inquirer dictionary may be considered as representing an operational explication of the scientist's theory or frame of reference. The investigator's questions represent the "rules" he develops within that theory for discriminating one kind of text from another. Contrary to hand content analysis procedures, both dictionary and question inquiries can be revised repeatedly without the necessity of recoding or repunching the original data.

Selecting tag categories. Since dictionary development is a crucial step in General Inquirer procedures, let us examine two different dictionaries currently in use. Figure 1 identifies the tag categories employed by our third general psycho-sociological dictionary[7] for analyzing texts of predominantly non-specialized vocabulary. Figure 2 gives the tag categories that Dr. Benjamin N. Colby*, an anthropologist, uses for studying themes in the folktales of different cultures.

Both Colby's dictionary and our own contain about 3500 entries each. Since the General Inquirer removes *s, es, ed, ly,* and *ing* suffix forms, the actual number of words that can be found in the dictionary is probably more than triple that number.

Our psycho-sociological dictionary makes a distinction between "first-order" and "second-order" tags: only one first-order tag may be used in categorizing a particular entry word, but one or several second-order tags may also be applied. All the tags in Colby's dictionary are first-order; each entry is associated with one tag and one tag only. Since multiple tagging of individual entry words can easily confuse later analyses, the first-order, second-order distinction should always be kept in mind.

* Dr. Colby is Associate Curator, Laboratory of Anthropology, Museum of New Mexico, Santa Fe, New Mexico. The dictionary and results referred to here are described in an article by Colby and Postal, currently in press in the journal, *Folklore.*

Figure 1. Third Psycho-Sociological Dictionary (Harvard)

FIRST ORDER TAGS

PERSONS

self—all pronoun references to the personal self (I, me, mine, myself)

selves—all pronoun references to the inclusive self (we, us, ours, etc.)

other—all non-sex-specific pronouns for other (you, yours, they, theirs, etc.)

male-role—all roles with specific male references

female-role—all roles with specific female references

neuter-role—all role names not connoting sex or occupations

job-role—all roles with clear occupational reference, theoretically open to both sexes

GROUPS

small-group—groups usually able to have face to face interaction

large-group—collectivities usually too large for face to face interaction

PHYSICAL OBJECTS

bodypart—parts of the body

food—articles or types of food

clothing—articles or types of clothing

tool—instrumental objects or artifacts of any kind (broader category than hand tools)

natural-object—objects not made by man (plants, animals, and minerals)

non-specif-obj—abstract references to objects (connoting intellectualization)

PHYSICAL QUALIFIERS

sensory-ref—smells, colors, tastes, etc.

time-ref—references to measurement of time

space-ref—references to spatial dimensions

quantity-ref—references to units and measures of quantity

ENVIRONMENTS

social-place—buildings and building parts; political, social, and economic locations

natural-world—geographical places, weather references and cosmic objects

CULTURE

ideal-value—culturally defined virtues, goals, valued conditions and activities

deviation—culturally devalued goals, conditions, and types of activities

action-norm—normative patterns of social behavior

message-form—names of communication media in a very broad sense, including art objects and money

thought-form—units and styles of reasoning

EMOTIONS

arousal—states of emotional excitement

urge—drive states

affection—indicants of close, positive, interpersonal relationships

pleasure—states of gratification

distress—states of despair, fear, guilt, shame, grief, failure, or indecision

anger—forms of aggressive expression

THOUGHT

sense—perception and awareness

think—cognitive processes

if—conditional words

equal—words denoting similarity

not—words denoting negation

cause—words denoting a cause-effect relation

defense-mechanism—standard psychological terms for defense mechanisms

EVALUATION

good—synonyms for good

bad—synonyms for bad

ought—words indicating a moral imperative

SOCIAL-EMOTIONAL ACTIONS

communicate—processes of transmission of meaning

approach—movement toward

guide—assistance and positive direction

control—limiting action

attack—destructive, hostile action

avoid—movement away from

follow—submissive action

IMPERSONAL ACTIONS

attempt—goal-directed activity, implying effort

work—task activity

get—obtaining, achieving action

possess—owning, consuming

expel—ejecting

SECOND ORDER TAGS

INSTITUTIONAL CONTEXTS—specification of the social context of roles and actions

academic
artistic
community
economic
family
legal
medical
military
political
recreational
religious
technological

STATUS CONNOTATIONS—male-, female-, neuter-, and job-role status implications

higher-status
peer-status
lower-status

PSYCHOLOGICAL THEMES

overstate—emphatic or exaggerative words, generally adjectives or adverbs (connotes a defensive style)
understate—words, generally adjectives or adverbs, connoting doubt or uncertainty (connotes a defensive style)
sign-strong—words connoting strength or capacity for action
sign-weak—words connoting weakness or incapacity for action
sign-accept—words implying interpersonal acceptance
sign-reject—words implying interpersonal rejection
male-theme—psychoanalytic symbols of masculinity
female-theme—psychoanalytic symbols of femininity
sex-theme—direct or indirect references to the sex act
ascend-theme—words associated with rising, falling, fire, and water, indicating concerns related to the Icarus complex
authority-theme—words connoting the existence or exercise of authority
danger-theme—words connoting alarm or concern with danger
death-theme—words connoting dying, end

Figure 2. Anthropological Dictionary (Colby)

1. Derived From Kluckhohn Value Categories
   Determine, Order, Indeterminate
   Dominate, Follow, Leader, Power, Agree
   Fame, Pride, Equality, Status
   Good, Evil, Suspicion
   Assist, Empathy, Guest, Selfish, Rivalry, Trick, Scapegoat
   Alone, Gregarious, Withdraw
   Choice, Individual, Group, Reject, Punish
   Chance, Curiosity, New, Protect, Caution
   Independence, Dependence, Ask
   Self-control, Over-indulge, Abstain
   Rational, Wise, Truth, Unknown, Naive, Foolish, Anger, Happy, Sad, Enthusiasm, Amuse, Dislike, Fear
   Tense, Relaxed
   Unique, General
   Quality, Quantity

2. Perception and Communication
   Aware, See, Smell, Talk, Taste, Sound, Quiet, Bright, Dark, Heavy, Hard, Hot, Cold, Color, Texture, Form
3. Space and Time
   Little, Big, High, Low, Narrow, Wide, Fast, Slow, Place, Time, Now, Old, Permanent, Future, Past, Building, Road
4. Self Identity
   Bodypart, Beauty, Bathe, Clothing, Healthy, Sick, Pain, Ugly, Ornament
5. Nature
   Sky, Earth, Air, Fire, Fluid, Dirt, Weather
6. Sex and Kinship
   Genital (Male and Female), Sex, Male, Malesymbol, Female, Marriage, Birth, Kinship, Kinship Affinal
7. Activities
   Hunt, Husbandry, Fish, Farm, Manufacture, Magic, Ritual, Exchange, Build, Repair
8. Miscellaneous Motif Groupings
   Able, Accomplish, Goal, Get, Keep, Increase, Want, Work, Plan
   Arrive, Leave, Move, Go Kinesthetic, Change
   Rest, Sleep, Difficult, Easy, Lazy, Unable, Fail
   Cover, Container, Complete, Empty, Full, Hole, Imperfect, Include, Uncover, Reveal, Release
   Boundary, Break, Cut, Pierce, Tear, Tie
   Anal, Oral, Food
   Death, Danger, Ruin, Lost, Secret, Prevent
   Money, Ownership

Our psycho-sociological dictionary uses 83 tags. With the exception of the categories "self", "selves", and "other", no tag category contains less than twenty words. Some second-order tags are used in categorizing over 300 words. While the number of tag categories is relatively small, further categorizing specificity is gained by considering those subgroups formed by common membership under both a first-order and a second-order tag. For example, those words matching both "job-role" and "legal" is a list in itself, quite different from those words matching both "job-role" and "academic" or "legal" and "ideal-value". These groups of words defined by first-order, second-order intersections can be directly specified in question retrieval procedures.

Inasmuch as sociological and psychological concerns can be separated, those tags with primary reference to the socio-cultural realm are listed in the left hand column of figure 1, those emphasizing psychological processes and themes are given in the right hand column. The sociological and anthropological concepts of roles, collectivities, actors, situations, values, norms, institutions, etc., are represented by tags under the headings entitled persons, groups, physical objects, qualifiers, environments, culture, institutions, and statuses. The psychological concepts of emotion and cognition, interpersonal and instrumental behavior are represented by the tags listed under the headings emotions, thought, evaluation, social-emotional actions, impersonal actions, and psychological themes.

If the distinction between the two columns is then combined with the distinction between first- and second-order tags, four distinct areas emerge. The first-order tags in the left hand column consist mainly of objects defined in sociological terms. The second-order tags in the left hand column refer to the social structure of society, specifically to institutional and status divisions. By contrast, the first-order tags on the right indicate basic psychological processes. The second-order tags on the right refer to some of the underlying psychological motivations of personality. Within each quadrant, the tags are arranged in sets. Wherever possible, the tags within a set are arranged in an order approximating a progression from the more personal and intimate to the more impersonal and objective.

The Colby anthropological dictionary uses 180 tags. The number of entry words for each tag ranges from 4 to 42. Many of the tag categories were developed as part of an attempt to validate cross-cultural ratings using the Clyde Kluckhohn value categories.[8] The tag categories have been built up through a substruction of the Kluckhohn binary categories (such as Good-Evil, Determinate-Indeterminate) into more precise units. In addition to the Kluckhohn categories, other motifs and themes which appeared to be important in the preliminary content analysis of folklore texts have been added. The tag categories are arranged under very general headings in figure 2 for clarity of presentation.

It should not be thought that all General Inquirer dictionaries must contain a large semi-comprehensive list of tag categories. For example, a General Inquirer dictionary[9] recently completed by the Stanford University project on International Conflict and Integration uses only tag categories relevant to the three major dimensions of Osgood's Semantic Differential[10]; namely, good-bad, strong-weak, and active-passive. Each of these dimensions has tag categories representing six levels of intensity, thus making a total of 18 different tags in all.

EXAMPLE APPLICATIONS

1. Comparing overt versus latent trends in folklore material.

The data. Approximately 12,000 words of folklore text from each of ten cultures were prepared for General Inquirer analysis. The ten cultures are: Kwakiutl, Egypt, Eskimo, India, China, Baiga, Russia, Kikuyu, Thailand, and Japan. Usually, such folktale materials have been gathered and recorded by missionaries and anthropologists. English translations, of course, have been used for purposes of the analysis.

The problem. Colby is interested in discovering themes which both characterize and distinguish cultures. More ambitiously, it is hoped that theme clusters may be discovered which provide insight into the way cultures or subcultures cognitively structure the world.

One problem of particular interest is the relationship between latent and overt themata in folklore material. In the results reported here, Colby compared the frequency of words denoting overt sexual activity with the frequency of words connoting words of latent sexual reference. Similarly, comparisons were made between overt and latent orality. Colby's latent sex index included references to possible symbols of the male sexual organ, to piercing or thrusting actions, and to possible symbols of the female sexual organ. The overt sex index included direct references to affection and sexual acts. The overt oral index consisted of words referring directly to the oral body zone and oral processes. The latent orality category is made up of what are thought to be symbolic references to ingestive states (full, empty, etc.); it is admittedly unclear, however, exactly how this latent oral category would relate to a latent anal syndrome.

The results. The rank ordering of the ten cultures on these four measures is shown in figure 3. As can be seen by inspection, the rank order for overt sex is inversely related to the rank order for latent sex (r = .55, prob. < .06), while the rank order for overt orality is posi-

Figure 3. Rank Order of Folktales on Four *Tag* Indices

Low Occurrence *High Occurrence*

Overt sex:
Eskimo, China, Kwakiutl, Kikuyu, India, Russia, Japan, Baiga, Egypt, Thailand

Latent sex:
Thailand, Baiga, Egypt, India, Kikuyu, Kwakiutl, Eskimo, China, Japan, Russia

Overt orality:
Kwakiutl, Japan, India, Egypt, Thailand, China, Kikuyu, Russia, Eskimo, Baiga

Latent orality:
Kwakiutl, Egypt, Thailand, Eskimo, Japan, India, Russia, China, Kikuyu, Baiga

tively related to the rank order for latent orality ($r = .66$, prob. $< .05$). With regard to sex, the implication seems to be that if open sexual references are inhibited in a culture, the sexual motives will nevertheless find indirect outlets. With regard to orality, several implications are suggested. Perhaps our latent category is not really latent. On the other hand, perhaps orality is not subject to processes of inhibition. Cultures low on orality are perhaps simply not making an issue of it, possibly because of their better nursing and weaning practices. Within a culture, sex is apparently always present in one form or another, whereas infant orality may be at least somewhat resolved.

2. Distinguishing real from simulated suicide notes.

The data. With the cooperation of the Coroner of Los Angeles County, Dr. Shneidman of the Los Angeles Suicide Prevention Center has collected 721 suicide notes from the court records of all recorded suicides for the ten year period 1945-1954. In each year, between 12 and 15 per cent of those committing suicide left notes. These notes come from both sexes (almost three males to every female), with the individuals ranging from twenty-five to fifty-nine years in age.

Simulated suicide notes were obtained from persons contacted in labor unions, fraternal groups, and the general community. Subjects were instructed as follows:

> "A study is being done on the prevention of suicide. For this, it is necessary to obtain many suicide notes written by normal people. For this reason, you are asked to write below, in your own words, the suicide note that you would write if you were going to take your own life. Make your note sound as real as you possibly can. Write what you think *you* would write if you were planning to commit suicide. Before you write the note, answer these two questions first:
>
> a) What method would you use to take your own life?
>
> b) To whom would you address the note you are writing?"

In order to keep the two groups homogeneous (and to emphasize whatever differences might exist in the notes), all sixty-six real and simulated notes were selected from those written by individuals who were male, Caucasian, Protestant, and native born. Each of the thirty-three simulated note writers was matched, man for man, with a real note writer who was not only of similar age (within five years), but also of the same occupational level.

The problem. Our research was twofold. As an academic exercise, we wanted to test whether a set of General Inquirer measures could be developed that together would effectively discriminate between real and simulated notes. Second, we were interested in what meaningful insights the Inquirer could offer as to the differences between these two groups.

The correct identification of which note in a pair is real and which is simulated is not a trivial challenge. Osgood[11] reports giving this same task to eight graduate students in psychology and finding that they could do no better than chance . One of us (Stone) gave this task to six members of a sophomore tutorial at Harvard, the students having no reading background particularly related to suicide. As a whole, the Harvard sophomores did better than chance, the mean being 66 per cent correct, the best performance being 75 per cent correct. Let us consider the 66 per cent as a base reference. Could the General Inquirer do better?

The results. The original analysis was done in February, 1962, using a predecessor of the Harvard Psycho-sociological Dictionary described above.[12] The procedure for building and testing a discriminate function was as follows: the actual source (i.e. real or simulated) of each of the first fifteen pairs of notes was revealed to us by Dr. Shneidman. These notes were then compared using the General Inquirer. Three factors were found to discriminate:

1) References to concrete things, persons, and places (higher for real notes).
2) Use of the actual word "love" in the text (higher for real notes).
3) Total number of references to processes of thought and decision (higher for simulated notes).

A very simple discriminate function was then developed: the score on the third measure was substracted from the sum of the scores of the first two measures. This index correctly discriminated thirteen of the fifteen pairs of notes.

This discriminate function was then applied to the remaining eighteen pairs of notes, with the members of the research team not knowing which of these were real and which were simulated. After the predictions were made, Dr. Shneidman was again consulted. Seventeen of the eighteen pairs of notes had been identified correctly. This figure is quite significant when compared with chance expectation, the performance of human judges, and most attempts of other investigators to analyze this same data (cf. review of literature in article by Ogilvie et al.).[13]

While constructing a successful discriminate function was a complex task, a complete General Inquirer analysis of the differences between the two kinds of notes was of a much larger order. After our Third Psycho-sociological Dictionary was finished, the suicide notes were again processed through the computer. At this point, of course, our research staff was no longer naive, so we could no longer make independent predictions on half of our data. Figures 4a and 4b show the differences in the actual number of times certain tags appeared in a specific syntax position in each of the two kinds of text.

Figure 4a. Raw Tag Counts Higher for Real Notes

FIRST ORDER WORDS

Tag Label	*Syntax Position*	*Number of Times Applied to Text*	
		Real	*Simulated*
other	attribute	17	6
male-role	subject	45	12
male-role	object	17	6
male-role	attribute	5	0
female-role	subject	200	67
female-role	object	137	35
female-role	attribute	21	5
tool	subject	5	0
tool	object	10	2
non-specif-obj	object	47	19
sensory-ref	object	6	0
quantity-ref	subject	27	10
social-place	object	27	4
affection	verb	51	18
bad	subject	9	2
bad	verb	9	0
communicate	attribute	14	3
attack	verb	17	2
attempt	leftover	23	9
get	verb	46	16
possess	verb	20	6

SECOND ORDER WORDS

Tag Label	*Syntax Position*	*Number of Times Applied to Text*	
		Real	*Simulated*
family	object	38	15
religious	subject	10	2
higher-status	subject	13	4
higher-status	object	11	1
higher-status	leftover	14	4
male-theme	(all)	18	1
sex-theme	subject	17	4
sex-theme	verb	46	13
sex-theme	object	8	2

Figure 4b. Raw Tag Counts Higher for Simulated Notes

FIRST ORDER WORDS

Tag Label	*Syntax Position*	*Number of Times Applied to Text*	
		Real	*Simulated*
selves	(all)	17	23
time-ref	object	3	10
natural-world	subject	4	11
natural-world	object	5	17
ideal-value	object	2	8
thought-form	subject	1	10
thought-form	object	3	10
distress	object	5	12
think	object	3	12

SECOND ORDER WORDS

Tag Label	*Syntax Position*	*Number of Times Applied to Text*	
		Real	*Simulated*
academic	object	5	12
sign-weak	subject	19	27
death-theme	(all)	54	62

In order to understand the criteria for selecting the tag labels which appear in figures 4a and 4b, we must first consider differences in the lengths of the notes. While shorter notes might be either real or simulated, most longer notes were real. This causes the thirty-three real notes to have a considerably greater combined length (total number of text words = 4112) than the simulated notes (total number of text words = 2542). Thus, unless the differences in raw frequency count for a particular tag label are considerably more or less than the differences in overall length, we are not impressed. Our criteria reflect this:

1) Select for figure 4a all cases where the total frequency of real suicide note tags is at least 2.5 times greater than the corresponding simulated note frequency, providing there is a minimum difference of five.
2) Select for figure 4b all cases where the total frequency of the simulated suicide notes is at least five counts more than the corresponding count for the real suicide notes.

Five is an arbitrary selection, reflecting what we feel to be a minimum difference to be of interest. Any difference in favor of the simulated notes is against odds (since the simulated notes are shorter), thus a simple difference of five is all that is required.

Figures 4a and 4b show a number of tag count differences which might be used in building discriminate functions. No information, however, is given on the distribution across notes. In actuality, some tag count differences are spread across many notes while others are concentrated in just a few. A tag may be highly discriminating for a few note pairs and be irrelevant in discriminating many other pairs. To account for this, we reason that there are multiple factors involved in suicide notes, some being a significant cue in one pair, others being a significant cue in another. If a cue is irrelevant to a particular pair of notes, then its prediction should be randomly determined. If, however, the cue is relevant, it should tend to predict in the correct direction. Adding together the different cues is like adding together signals in a background of noise. In some cases, one cue will be the signal while several others are noise; in another case, several of the other cues may serve as signals while the first cue functions as noise. In a similar vein, we can afford to have a limited amount of "mistagging" in our procedures, providing it acts like random noise in our analyses. If enough cues are added together, the signals usually emerge from the noise and a correct prediction can be made.

The syntax positions shown in figures 4a and 4b are "subject", "verb", "object", "attribute", "leftover", and "all" (for all syntax positions combined). These syntax marking procedures are described in detail elsewhere[1]. The terms "subject", "verb", and "object" are used in a sense similar to, but slightly broader than, the meanings you learned in grammar school.[11] "Attribute" indicates those words in a phrase indicating the source of a statement, such as, "He says . . .", "She knows that . . .", "It is true that . . .", etc. "Leftover" refers to counts for syntax positions not listed here and for words lacking syntax marks. As we shall see, reporting tag counts for each major syntax position separately is a useful aid in analyzing the data. Since the counts for each syntax position are based on only a few words per sentence, a separate analysis of each syntax position helps keep counts from being determined by multiple occurrences in just a few sentences.

The information in figures 4a and 4b can be combined with retrieval procedures to gain more information about the text. For example, the tag "possess" is shown on the bottom of the right hand column of figure 4a as being used in the verb position more frequently in real notes than in simulated notes. Who is seen as doing the possessing? We see further up in this same column that there are several likely possibilities. Male-role and female-role, for example, are both relatively high in the subject position for real suicide notes. If we make the retrievals "male-role/subject—possess/verb" and "female-role/subject—possess/verb", we find that it is the sentences containing female rather than male references as subject that are causing high counts. The retrieval for the question "male-role/subject—possess/verb" consists of only one sentence and this from the text of a simulated note:

note 15, sentence #11

GOD BLESS AND KEEP YOU BOTH.

The retrieval for "female-role/subject—possess/verb" brings back seven sentences from the real notes and no sentences from the simulated notes. The seven sentences are:

note 4, sentence #21
SHE KEPT AFTER ME.

note 6, sentence #18
BUT GAITY YOU (WOMAN) SAVED FOR STRANGERS.

note 10, sentence #7
AND SHE WOULD STAY WITH ME.

note 23, sentence #7
TOO BAD YOU (WOMAN) JUST KEPT EVERYTHING INSIDE YOU (WOMAN).

note 24, sentence #28
YOU (WOMAN) HAVE EVEN MORE THAN YOU (WOMAN) HOPED FOR.

note 26, sentence #10
I HOPE YOU (WOMAN) HAVE ALL THE LUCK IN THE WORLD.

note 29, sentence #5
(YOU [WOMAN]) KEEP EVERYTHING QUIET AS POSSIBLE.

These retrieved sentences tend to picture women as being rather powerful, the ultimate deciders of what is or is not shared with, or given to, the writer.

Note in these retrieved sentences the importance of text editing. Part of the optional preparation procedure involves identifying proper names and ambiguous pronouns. Five of these seven sentences would have been missed if the sex of the second person pronoun were not identified. Syntax marking is used to separate the verb "have" as a main verb from "have" used as an auxiliary. Only "have" used as a main verb is tagged with the label "possess".

Many of the other sentences having "female-role" as subject are concerned with giving instructions in both the real and the simulated notes. In an analysis by Ogilvie et al.,[13] using our earlier dictionary, it was found that these requests in the real suicide notes were rather specific, such as, "YOU (WOMAN) TELL MY FOLKS," or "YOU (WOMAN) PLEASE TAKE CARE OF MY BILLS," while the requests in the simulated notes were much more vague, e.g. "YOU (WOMAN) FIND A NEW LIFE FOR YOURSELF."

Question retrievals often help to locate the cause of a particular tag count. In the left hand column of figure 4b, for example, we noticed that the tag "natural-world" was high for simulated notes in both subject and object positions. In looking at a retrieval printout of this category, we noticed that many of the tag counts for the simulated notes were caused by the text word "life", particularly as an object in conjunction with the tag "self" as subject. While life itself is a part of nature, it certainly differs from (and is admittedly somewhat misplaced with) other entry words in this category, such as those pertaining to weather and cosmic objects. A special retrieval was made for all those sentences that combined the tag "self" as subject with the appearance of the actual text word "life" as an object. Eleven sentences were retrieved from the text of simulated notes, only three from the real notes. Given the fact the real text is considerably longer, this is quite contrary to chance expectation. The retrievals are well distributed over the entire text. Both real and simulated note retrievals yield messages quite similar in content: "I CANNOT FIND MY PLACE IN LIFE." "BUT I JUST CANNOT STAND LIFE ANY LONGER." "YOU SEE, I KNOW NOW THAT I CAN NEVER HOPE TO REALLY MAKE A SUCCESS OF LIFE." "BUT I CAN NOT SEEM TO STAND LIFE THIS WAY." "I HAVE NOT MADE LIFE SEEM WORTHWHILE." "AND I CAN SEE NO USE IN PROLONGING IT (LIFE)." "IN A FEW MINUTES I WILL TAKE MY LIFE," etc. All but the first of these sentences are from the simulated notes.

The results presented here are meant only as examples of the kinds of analyses that can be done using General Inquirer operations. The tag categories listed in figure 4 are not the only ones to be used in making analyses. Two categories, each of which may not differentiate the two texts by itself, may very well serve to differentiate when combined in a retrieval question. We have already written two articles[13, 14] describing our work with these notes. A third article probably will be written this summer. Suffice it to say here, in summary, that the notes can be differentiated by the General Inquirer.

Our analyses find real suicide notes characterized by references to concrete events and interpersonal relationships (the role of women as powerful and denying being of particular significance), and simulated notes characterized by their abstract intellectualization, often amounting to a reflection on the relative merits of life and death.

AUTOMATIC THEME ANALYSIS

One of the problems frequently facing the General Inquirer user is the construction of question sets that tap more than a small fraction of the sentences in the text. For example, an investigator may be able to develop questions that are very powerful in discriminating sentences of source *A* from those of source *B*, but the number of sentences used in making such discriminations may involve a total of as little as twenty per cent of the entire text. What about the rest of the text? May it not also have some discriminating features?

One possibility is that the remaining text does contain features that would permit discrimination between the two sources, but that the investigator must ask different questions in order to find them. Perhaps the investigator's theory is orthogonal to the real differences lying within the texts. Another possibility is that there are real differences between the texts, but that the dictionary itself is not relevant. If this is the case, then there is no better alternative than to develop better dictionaries. Finally, there remains the possibility that there are no differences between the two texts. If this is the case, we are stuck. However, as we shall see, the actual state of "no difference at all" in such a multiple descriptive instrument as the General Inquirer is quite rare.

From our experience in using the General Inquirer, we came to feel that the main difficulty in developing rules that would discriminate a greater percentage of the text was due to our rule building procedures, not to any inadequacies of our dictionaries or the lack of real differences in the text itself. Dr. Hunt, who was then at Yale, suggested that perhaps we could use computer concept formation procedures to help find more discriminating rules. He offered his own Concept Learner[3] as a computer system for putting this into effect. Probably many persons in this audience remember hearing the Hunt-Hovland predecessor of this system described at the 1961 Western Joint Computer Conference.[2] The current Concept Learner (CL-2) is very large and has application to a wide variety of problems. There was no doubt that CL-2 could do at least an "interesting" job of building rules for discriminating between two text sources.

One source of considerable humor in the last few years has been the characterization of computers as having semi-human qualities. While the idea of two computing machines marrying and having baby computers is perhaps best left to fantasy, the merging of two large scale computer program systems to produce a new program system with an identity of its own is indeed within the realm of possibility. What follows is the story of such a romance.

As with any large scale affair, there were some technical difficulties. In this case, our two principals, the General Inquirer and the Concept Learner, did not speak the same language. The General Inquirer is programmed entirely in COMIT, a computer language developed by Yngve and his mechanical translation group at M.I.T.[15] The Concept Learner is programmed entirely in IPL-V, a computer language that came from Carnegie Tech and the RAND Corporation.[16] Fortunately, the General Inquirer and the Concept Learner had two bi-lingual friends, namely Hunt and myself (Stone), to act as interpreters. A programmed interlingua was developed for handling translation problems and the merged operation was on its way.

The resulting procedure is this: the General Inquirer gives to the Concept Learner (via our interlingua) a list of tags that have been applied to each sentence of both texts *A* and *B*. While the original text words are dropped, care is taken to note the syntactic position that was associated with each tagging operation. Each sentence is thus replaced with a list of syntax marked tags (hereafter called "labels"). Given sentences described in this form, the task for the Concept Learner is to find a near minimum set of "rules" for distinguishing as many sentences as possible of document *A* from those of document *B*.

Rather than have to consider all the complexities and interrelationships of actual language, the Concept Learner is provided by the General Inquirer with a description of each sentence consisting of a list of tags chosen from the 83 different categories, each tag marked with one of the five different syntax positions. There is thus a finite resource of 415 different labels from which the dozen or so symbols used in describing any particular sentence must be selected. The number of combinations of 415 labels, taken a dozen or so at a time, is very large, especially when any one label may be used more than once in a sentence. It is not surprising that we find it very rare for two or more sentences to have identical descriptive lists. So long as there remain differences in these descriptive lists, we have grist for building discriminative functions.

The Concept Learner looks at all the sentences in each document to see if there are one or more labels common to all sentences in one document that are not found in any sentences of the other document. For example, if *all* the sentences in document *A* were retrieved by these two labels:

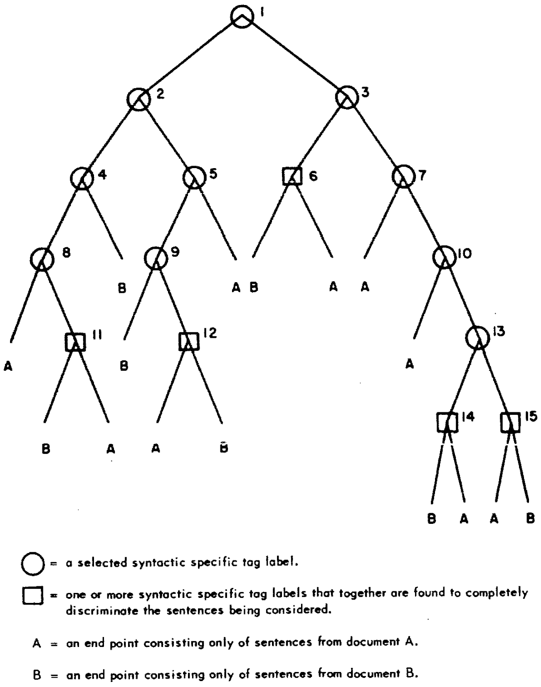
1) The tag "family" in the syntactic position of subject.
2) The tag "control" in the syntactic position of verb.

but *none* of the sentences in document *B* matched both these specifications, then this single question, by itself, could be considered as completely discriminating.

Usually no single question can successfully serve as an "all-none" test for discriminating all the sentences in one document from all the sentences in another. Thus, it becomes necessary to develop a discriminating procedure that uses a number of different questions. The tree procedure is one means for doing this.

The basic tree building strategy in CL-2 is based on the following heuristic: if there is not a single label or combination of labels that completely discriminates the sentences in document *A* from those in document *B*, the Concept Learner takes whichever document has the fewest sentences and finds which label occurs in the most of these sentences. The occurrence of this label is then used as a test to divide the sentences in documents *A* and *B* into two subsets each, depending on whether the label is present (sets A_{1+}, B_{1+}) or absent (sets A_{1-}, B_{1-}). The search for a successful all-none discriminating test can now be applied separately to each of these new subgroup pairs. First, subgroup pair A_{1+}, B_{1+} is tested. If a successful all-none discrimination is made on this subgroup, the computer can immediately turn its attention to finding a successful all-none test for subgroup A_{1-}, B_{1-}. If no successful all-none test is found, the subdivision process continues until either a sub-subgroup is found where an all-none test does apply or one of the document sources runs out of sentences.

Figure 5 shows an example of this subdivision process. A branch to the left represents the sentences that successfully matched the specifications at this node. A branch to the right represents the sentences that failed to match this specification. In actual operation, the tree keeps subdividing to the left until a complete discrimination is made. It then works on the A_{n-}, B_{n-} subgroups, starting from the bottom subgroup (i.e. the right branch of node #8) and working back up.


1
2
3
4
5
6
7
8
9
10
B
A B
A A
11
12
A
B
A
13
B
A
A
B
14
15
B A A B
○ = a selected syntactic specific tag label.
□ = one or more syntactic specific tag labels that together are found to completely discriminate the sentences being considered.
A = an end point consisting only of sentences from document A.
B = an end point consisting only of sentences from document B.

Some results using the Concept Learner on text analysis problems. There are several criteria that are useful for evaluating the success of the Concept Learner. One is the simplicity of the tree. Obviously, if the total number of nodes in the tree almost equals the number of sentences in both documents combined, we have not done much towards finding trends in our data. A second test is how well a developed tree can identify correctly the source of new sentences. To test this, we divide our data approximately in half (as we did in the suicide note study) and use only half the data to develop a tree. The tree is then used to classify the other half of the data, the percentage of correct classifications being our evaluation index.

So far, the Concept Learner has been used on about a half dozen automatic theme analysis problems. Identifying reliable differences where each source is the work of just one author turns out to be much easier than identifying reliable differences where each source is the work of a group of authors. Most of the trees developed on multiple author sources have been quite complicated, some containing upwards of sixty test nodes. Usually, these more complicated trees are not very good at classifying additional test data correctly.

The suicide note problem, of course, is one where each source contains multiple authors. Partly because of this, the Concept Learner yields a rather complicated tree that has less success in discriminating real from simulated notes than our regular man-machine General Inquirer procedures reported above. In one respect, this lack of complete success with all machine procedures is somewhat of a relief. After all, we do not want to create technological unemployment for ourselves as psychologists.

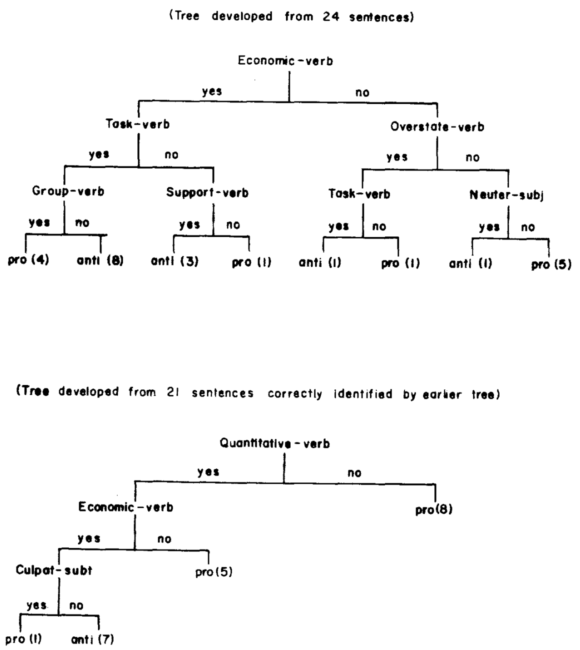
As a less complex example, let us consider a study involving limited amounts of data with only one author to each source. The data here consists of four documents circulated in connection with the 1962 California state election.

California is famous for requiring its citizens to act as legislators. The typical California election ballot presents the voter with twenty to thirty quite complex issues. To aid him in decision making, the Secretary of State distributes a booklet describing the various propositions on the ballot. This booklet contains brief arguments by proponents and opponents of each measure.

In 1962, an unusual event occurred. Substantially the same proposal appeared on the ballot twice. Proposition 1, which arose in the State Assembly, was a proposal which would have authorized an increase in the pay of state legislators to $11,000 a year. Proposition 17 was a proposal by the State Senate to raise salaries to $10,500. Different paragraphs were included in the booklet urging and opposing each measure. The arguments were quite close in each case. Both opposing arguments were written by a California property owners association, the supporting arguments by committees of the legislature. (As might be expected, the voters apparently perceived the similarity between the two propositions; both measures were defeated at the polls.)

Twenty-four sentences were selected from the arguments for and against Proposition 1, and twenty-eight sentences from the arguments for and against Proposition 17. Sentences which obviously indicated the source, such as "Vote no . . .," were excluded.

The sentences for and against Proposition 1 were processed through the General Inquirer and then analyzed by the Concept Learner. The result of the analysis is the tree shown in figure 6a. The original twenty-four sentences break down into eight groups, each group corresponding to an end point in the tree. Four of these groups collectively contain all the pro statements; the other four groups contain all the opposing statements. Thus, if the inquiry is: "economic/verb, work/verb, and group/verb,"


(Tree developed from 24 sentences)

(Tree developed from 21 sentences correctly identified by earlier tree) there will be only four sentences in the text that match all these specifications, and they will all be in support of the proposition. However, the discrimination rule "economic/verb, work/verb, and *not* group/verb" refers to a set of eight sentences, all of them against the issue. From the point of view of making General Inquirer retrievals, the entire tree structure can be converted directly into "question sets"; each branch of the tree defines a separate question.

Clearly, the amount of data reduction in figure 6a is insufficient. Four of the eight end points refer to only one sentence. These probably represent specialized questions that are very unlikely to be of discriminative value in future analyses. Probably more powerful trees can be developed to represent the differentiating main trends in the data. Note, for example, that "economic/verb" by itself is able to distinguish eleven of the thirteen anti sentences.

Nevertheless, this tree does fairly well when used to classify the twenty-eight sentences from Proposition 17. Twenty-one of the twenty-eight sentences were classified correctly. If these twenty-one correctly classified sentences are given to the Concept Learner, a new, much simpler tree can be grown, as shown in figure 6b. Basically, the sentences against the propositions are concerned with "economic/verb, quantitative/verb" aspects of the issue. Non-economic quantity references are in favor of the issue, as are sentences not mentioning quantity references in the verb at all. Those arguing against the bill are apparently preoccupied with its economic costs; those supporting the bill tend to focus on its other features. Granted, most California citizens could have told us this without our having to go to so much trouble. Yet our machine has done thus via explicit procedures, where our average citizen might be a bit hard pressed to explain his exact cognitive reasoning. Maybe we can use one to help understand the other.

As we have seen, the Concept Learner builds trees using two main heuristics: an "all-none" test for terminal nodes and special non-terminal procedures for dividing text into subsets. There are many other procedures that might also be explored. We have decided to expand the Concept Learner to handle these other possibilities. Given the inordinate proportions of computer time consumed by our interlingua routines, it was decided to use this opportunity to reprogram the Concept Learner, plus the new procedures, in the COMIT computer language. COMIT has proven to be well suited to the task. The program has been designed and written at Harvard by Mr. Marshall Smith and is currently in the debugging stage. Hopefully, it will be running at the time of this meeting.

The principal difference between the new program and CL-2 is the number of tests available at a given node. By using flexible dispatching procedures, the investigator can select tests from the following list and order them in any way he wants. All tests offer minimum cutoff parameters which are again adjustable by the investigator. The list of tests is as follows:

1. *All-none test based on a single label.* Looks for a label that characterizes all the sentences in source *A* but none in source *B*. This failing, it looks for a label occurring in all the sentences of source *B*, but none in source *A*.
2. *Some-none test based on a single label.* Looks for the label that both occurs in the largest number of sentences in source *A* that is not represented in source *B*. A similar search is made of labels in source *B* that do not occur in source *A*.
3. *Many-few test based on a single label.* A search is made for that label which has the largest absolute difference in the number of occurrences in source *A* versus source *B*. If the number of "few" sentences is below a certain minimum specified by the investigator, they are thrown away and the left hand side of the node is marked as being terminal. If the number of "few" sentences is above this minimum, the left hand side is considered non-terminal, and further subdivision occurs.
4. *Hunt non-terminal heuristic.* The computer determines which source contains the fewest sentences at this node and finds the label that occurs in the largest number of these sentences.

By basing all our tests on a single label, it is possible to execute tests quickly, using matrix procedures rather than list commonality searches. The cost is that we cannot then develop single node discriminations based on a conjunctive combination of labels. While this is essential for other Concept Learner problems, it is not necessary for automatic theme analysis when some-none or many-few tests are also provided.

One addition planned for the near future is a "look ahead and evaluate the various possible outcomes" routine that will allow the investigator to direct the tree building procedures according to his own theoretical interests. Since it would be inconceivable for the machine to search ahead and try out all possible outcomes, it will look for a list of suggestions supplied by the card reader. All these suggestions will then be evaluated by the computer in terms of the extent to which they lead to certain desirable tree properties. The most highly evaluated suggestions will then be used in building the tree.

FINAL REMARKS

While most of our research examples have been given mainly to illustrate procedures, we would not want to leave the impression that we are blind to the many possible General Inquirer applications. The General Inquirer is currently being employed by different investigators on a variety of theoretical and applied problems, both serious and sometimes somewhat amusing. Data already under study include small group discussion procedures, diplomatic notes exchanged between countries, personality differences between northern and southern negroes, congressional subcommittee testimony of different lobby organizations, delusional language of schizophrenics, college applications, descriptions of magico-religious role differentiation in primitive societies, differences in attitudes toward friendship in large and small town environments, peace corps field reports, comparison of police chiefs and probation officers regarding their attitudes toward the juvenile delinquent, the letters written by the "Three Faces of Eve" as a study in multiple personality, cross-national comparison of college students' plans and expectations for the future, and thematic changes in popular song lyrics during the last 30 years. While we cannot discuss so many studies in detail here, the results so far have been very encouraging. Pleased with our initial results, we hold optimistic expectations for the future.

REFERENCES

1. STONE, P. J., BALES, R. F., NAMENWIRTH, J. Z. and OGILVIE, D. M. The General Inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7, 1962, 484-498.
2. HUNT, E. B. The development of decision trees in concept learning: model and basic results. Working paper number 6, Western Management Science Institute, U.C.L.A. April, 1962.
3. HUNT, E. B. *Concept Learning: an Information Processing Problem.* New York: Wiley, 1962.
4. STONE, P. J. and HUNT, E. B. The General Inquirer extended: automatic theme analysis using tree building procedures. *IFIP Proceedings,* Munich, 1962.
5. BERELSON, B. Content Analysis, in Lindzey, G. (ed.) *Handbook of Social Psychology.* Cambridge: Addison-Wesley, 1954.
6. MOSTELLER, F. and WALLACE, D. L. Inference in an authorship problem: a comparative study of discrimination methods applied to the authorship of the Federalist papers. Department of Statistics, Harvard, 1962.
7. MCPHERSON, W., DUNPHY, D., BALES, R. F., STONE, P. and OGILVIE, D. M. A Revised Psychological and Sociological Dictionary for the General Inquirer. Dittoed paper (two volumes) Laboratory of Social Relations, Harvard University, December, 1962.
8. KLUCKHOHN, C. The scientific study of values. University of Toronto Installation Lectures, 1958.
9. HOLSTI, O. R. Computer content analysis. Working papers numbers 1, 2, and 3. Stanford Studies in International Conflict and Integration, 1963.
10. OSGOOD, C. E., SUCI, G. J., and TANNENBAUM, P. H. *The Measurement of Meaning.* Urbana: The University of Illinois Press, 1957.
11. OSGOOD, C. E. The effects of motivation on style of encoding. In Sebeok, T. (ed.) *Style in Language.* New York: Wiley, 1960.
12. BALES, R. F. and STONE, P. J. A general psycho-sociological dictionary for the Gen-

eral Inquirer. Laboratory of Social Relations, Harvard University, 1961. Ditto (also briefly described in Stone and Bales, *et al.*)[1]

13. OGILVIE, D. M., DUNPHY, D. C., SMITH, C., STONE, P. J., with SHNEIDMAN, E., and FARBEROW, N. Some characteristics of genuine versus simulated suicide notes as analyzed by a computer system called the General Inquirer. Laboratory of Social Relations, Harvard University, August, 1962. Ditto.

14. STONE, P. J., OGILVIE, D. M., and DUNPHY, D. C. Distinguishing real from simulated suicide notes using General Inquirer procedures. Paper read at the joint annual meeting of the American College of Neuropsychopharmacology, Washington, D. C., January 25, 1963.

15. YNGVE, V. *COMIT Reference Manual.* Massachusetts Institute of Technology Press, 1962.

16. NEWELL, A. (ed.) *Information Processing Language V Manual.* New Jersey: Prentice Hall, 1962.