Progress in Communication Sciences

Volume XII

edited by

William D. Richards, Jr.
Simon Fraser University

George A. Barnett
State University of New York at Buffalo

Ablex Publishing Corporation
Norwood, New Jersey 07648

1993

9 Network Analysis of Message Content

James A. Danowski
Department of Communication
University of Illinois at Chicago

This chapter presents a theoretical model of messages and communication processes. It also describes a way to create optimal persuasive messages from natural language texts. The model avoids the limitations of categorical content analysis, both manual and computerized. Network analysis provides a relational paradigm for statistically based message rsearch; it performs a highly automated *qualitative* analysis through the use of *quantitative* procedures. Several examples illustrate applications of word-network theory and method to different problems.

I. CONCEPTUALIZING SEMANTIC NETWORKS

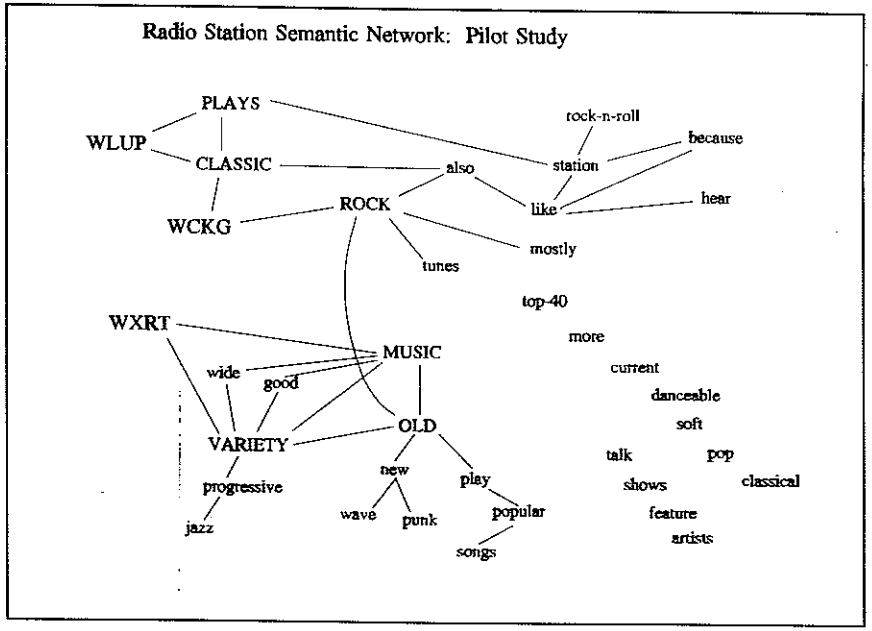
Imagine a large group of people using the same language over time. Assume that the full text of their messages is available to you in natural language form. How would you come to some representation of what they are talking about?

Your first thought may be to use traditional content analysis methods that categorize text, either manual procedures or computerized ones like the General Inquirer. These, however, are relatively crude. They merely assign message elements to a limited number of nominal categories.

Instead of categorizing messages, with a network perspective you can capture the relationships among words within the messages. Defining word-pair link strength as the number of times each word occurs with another, every possible word pair has an occurrence distribution, whose values can range from zero on up. This ratio scale of measurement allows you to use sophisticated statistical tools. Some of these enable you to map the structure of the word network. They identify word groups, or clusters, and quantify the structure of the network at different levels. Using these word-pair data as input to network analysis tools, you map the language region. On the map, instead of cities, the nodes you see are words. Rather than roads, there are links among words. Figure 1 shows an example of a word-network map from descriptions of listeners' favorite radio stations.

Travelling through the word network are fleets of social objects. These communication vehicles are the concepts, ideas, or physical things that people linguistically describe. As they link words to these vehicles in the course of their everyday informal and formal communication, this propels them through the network. Sometimes these movements are unplanned. At other times, organizations try to manage vehicular traffic. By means of optimal messages, they try to steer vehicles away from certain words or to induce traffic to flow in particular word-link areas (Danowski, 1990).

Mathematically-based procedures have been developed to create optimal messages. These are constructed through systematic analysis of the paths connecting word nodes of



interest. The procedures identify the optimal association network across the aggregate social community. The underlying assumption is that stimulating associations across it is more effective as the shortest efficient sequence of words is selected for the message. This is because people process strings of words linearly over time, encoding and decoding them in sequences. Furthermore, the triggering of associations to words in context takes cognitive time. The most effective message, therefore, optimizes the association networks in the receivers' minds as they read or hear the message.

Based on the preceding conceptual platform, the focus of this chapter is on text analysis and text generation. As such, it lies in the content analysis domain. Yet, it is one that is large and diverse. To distinguish the network approach to content analysis explicated here, the next section contrasts it with some aspects of traditional content analysis.

II. A CRITIQUE OF TRADITIONAL CONTENT ANALYSIS APPROACHES

Content analysis has been a central methodological tool for testing communication hypotheses involving message variables (Krippendorff, 1980). Nevertheless, traditional

content analysis has some limitations:

1. It is primarily categorical in orientation. Lewin (1931) criticized categorical measurement as excessively "Aristotelian" and limited. He argued instead for a "Galilean" approach to measurement that focuses on continuous variables, on forces and associations, not on categories and differences.

2. Even though theoretically driven, categorical content analysis is quite conceptually crude. It throws away a lot of information as it forces content elements into a relatively small set of mutually exclusive categories. Both manual and computerized content analysis procedures are primarily categorical. Most computer approaches (Weber, 1990) assign textual units to categories. They look up in a hand-built dictionary file each word's category assignments, and then tabulate the category percentages. Both the General Inquirer (Stone, et al., 1966) and General Inquirer III.

3. Categorical content analysis is relatively expensive. In the case of manual content analysis, there is a high labor cost. It requires a relatively sophisticated symbolic analyst. And, such coding moves slowly, even for experienced analysts. Computerized categorical analysis is also costly in terms of computing resources. It requires large dictionaries containing all possibly occurring content elements. At the same time, building these reference files takes a lot of manual human analyst time. Further-more, once built, the sheer size of the reference dictionary requires extensive comput-ing resources. Most categorizer programs run only on mainframes, not on PCs.

4. Traditional content analysis looks only at the categories. Relational content analysis, which focuses on patterns of relationships rather than on categories, is also costly. Such pattern recognition requires even more extensive manual time and talent. If computerized, however, the limits are not as severe as with categorical analysis. The pattern recognition power of computers and software can be marshalled, instead of allowed to impose mere brute force.

5. Linguistics-based perspectives take a parsing approach. They assign units of language to grammatical categories. Similar to the general categorical approach, it is at a more micro-level. Furthermore, it derives from tightly constructed grammatical theories, thus constituting a highly top-down parsing of text. These perspectives apply grammatical theory to classify portions of it into different constructions; for example, parts of speech. These approaches stand out in their striving for consistency, completeness, and conceptual correctness. Nevertheless, they often treat only toy data, analyzing single sentences or phrases to exercise the scholarly issue at hand. Because of this, the linguistics-based approaches are usually quite limited in external validity. They are not often applied to large, representative samples of text (Wilks, 1989).

6. Besides the linguistic and the statistical approaches to content analysis, there is the critical perspective. It is ideologically based: to maintain oppression, the dominant capitalist coalition controls the language of media content. Because of the politically motivated nature of this scholarship, it is highly selective in its approach to content analysis (Hall, 1988). It purposively select language evidence that supports its rhetorical objectives, and argues for the use of politically correct language.

7. Linguistic semantics is another approach to content analysis. Its primary applications are in artificial intelligence circles. Typically, humans must manually code each word in a large dictionary into semantic categories. These include: actor, action, patient, object, time, space, process, and event (Wilks, 1989). Once the coders have laboriously hand-built the dictionary, the software then runs largely in an automated mode. It parses incoming natural language sentences and represents them in terms of relation-ships among these categories.

   The main limitation of linguistic semantics has been its focus on deterministic, micro-level language units. Like linguistics of a more grammatical bent, linguistic semantics often deals with rather toy-like data sets. The researcher constructs thesm to create fast, workable operationalizations. Unfortunately, once the proof of concept has been made, researchers may be less likely to drive the procedures into the turbulent domain of large-scale, real-time, natural language processing.

8. Regardless of approach -- be it linguistic-grammatical, linguistic-semantic, critical, or statistical -- another serious limitation to content analysis has been in the availability of machine-readable textual collections, or corpora. As a result, pitting one method against another on the same externally valid playing field has been difficult. Nevertheless, there is promise for overcoming this limitation. An international Text Encoding Initiative has been working now for several years, creating standards for preparation of machine readable text (Hockey, Burnard & Sperberg-McQueen, 1991).

9. A final limitation of prior work connects with traditional notions about qualitative versus quantitative approaches to gathering and representing information. The old view of qualitative research encompassed any sort in which the analyst did not count or measure anything. On the other hand, the quantitative research label was applied any time there was any numerical analysis. Today, qualitative research can be highly statistical and numerically based. The main difference between quantitative and qualitative analysis lies in the discrete quality of the latter's dependent variables. In quantitative studies they are continuous. Another type of qualitative analysis focuses on relationships among the units studied, not solely on their individual attributes. Because of the traditional view, scholars used to think that messages and meanings were treatable only without numbers. However, network approaches to content analysis are not only highly qualitative, but also simultaneously quantitative. For example, statistical, network-based content analysis is primarily a qualitative procedure.

III. NETWORK CONTENT ANALYSIS

With network content analysis, the message scientist can more effectively design research to serve a range of purposes: 1) description, 2) prediction/explanation, and 3) control. More specifically, the word-network approach can be used for:

a) Description of communication content,
b) Theory development,
c) Hypothesis testing,
d) Identification of social units and problems,
e) Evaluation, selection, and placement of units,
f) Message creation, and
g) Effects tracking

Network approaches to content analysis have not frequently appeared in the literature. What examples exist use manual coding of networks of actors represented in the content. Each time an actor appears in the text with another, the pair receives a score. After all content has been coded, then network analysis is performed using actors as nodes. Danowski (1988) coded a set of news stories for relations among all organizations appearing across them, treating them as nodes in defining a focal organization's interorganizational field. Alexander & Danowski (1990) network analyzed individual actors mentioned in Cicero's letters more than two thousand years ago. While useful for some purposes, manual actor-network coding is not the type of network content analysis of central concern in this chapter.

In contrast, here the focus is on the range of word cooccurrences across texts (Danowski & Martin, 1979; Danowski, 1980a, 1980b; 1982), not just actor names. Although our initial word cooccurrences pilot studies used manual coding, now they are automated. The approach can be thought of as using a window, 'n' word positions wide, which it slides through the text, counting and aggregating all word pairs within the window. Then, the word pairs are network analyzed. This identifies an aggregate structure for the whole set of words. It finds clusters or groups of words that frequently occur together. It also finds liaison words, linking groups together. In addition, it computes quantitative measures of the structural properties of the overall network, as well as of subareas and of individual nodes.

Once the word-network analysis is completed, the analyst can link the textual representations to other information. Adding such "side data" enables tests of hypotheses about relationships of word network variables to others kinds. For example, message variables can be tied to markets. One example is to correlate word pairs with financial performance data, such as stock price-to-earnings ratios. The associations between the words and the financial valuation metric provide a way to measure the predicted value of any financial message (Danowski, 1991b). A second example uses market share, instead

of financial performance. Here, word pairs that the audience uses to describe the marketing vehicles are correlated with their market share: audience ratings data indexing market share are correlated with word-pair frequencies that audience members use to describe the radio stations (Danowski, 1991a).

IV. STEPS IN WORD-NETWORK ANALYSIS

Figure 2 on the next page summarizes the primary procedures in this author's approach to word-network analysis of natural-language. WORDLINK is a computer program we developed to process text files into word-pair records.

1) Assemble textual corpus

The analyst assembles the text in regular ASCI files. First it is useful to run a spellchecker and correct errors; even commercial data base records contain them. With respect to text sample sizes, general statistical guidelines suggest that the minimum number of cases should be 30, with numbers in the neighborhood of 300 ideal. A variety of different kinds of text have been analyzed:

a) electronic mail transcripts;

b) news stories;

c) advertising copy;

d) book excerpts;

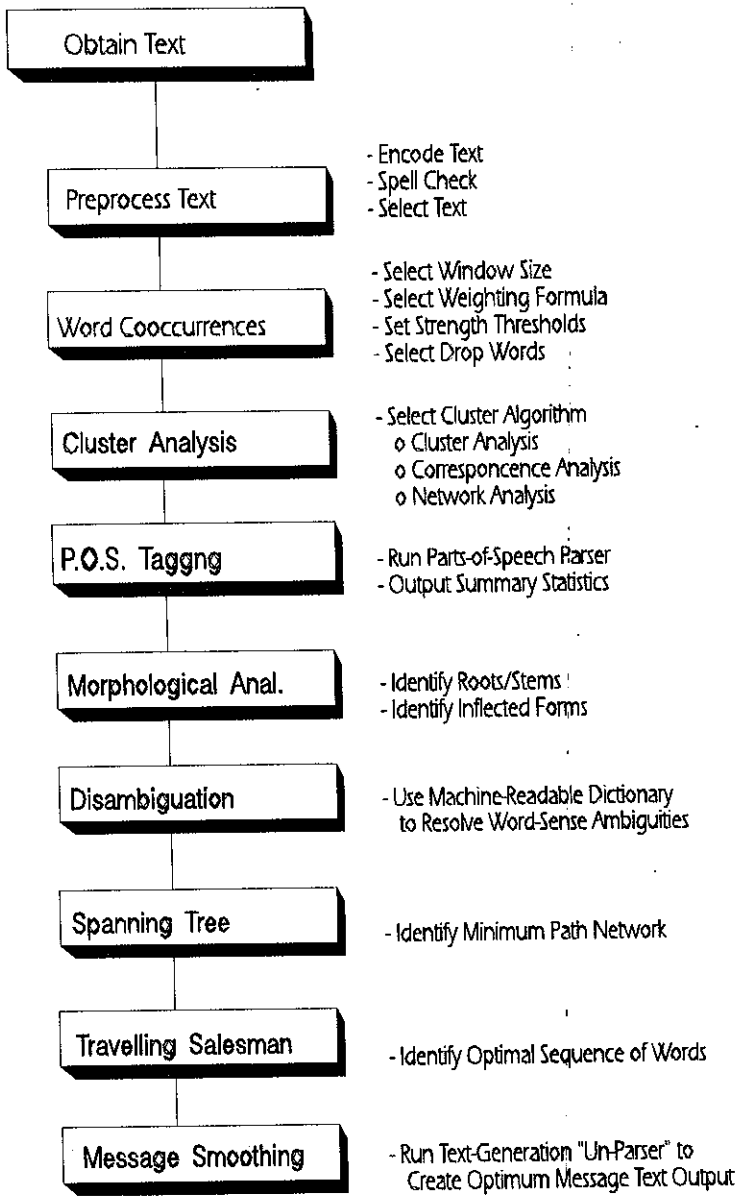
e) focus group transcripts;

f) verbatim responses to open-end survey questions, including:
   - self-administered paper-and-pencil questionnaire forms, and
   - computer-assisted telephone interviewing, in which the protocol appears on the interviewer's computer screen. As the interviewer asks questions, he or she types verbatim responses, which are stored on disk.

Table 1 provides an example of input text. These are a sampling of verbatim answers to an open-ended survey question asking UIC students to describe their favorite radio station. If different groups are to be compared, then the analyst creates files containing the record identification numbers for each group. The WORDLINK software then selects records, running a parallel analysis for each group.

2) Do word frequency counts

The next step is to identify the unique words and their frequencies of occurrence. The utility of this includes identifying the total number of words for use as a denominator.



should the analyst want to compute proportional frequencies. Table 2 shows a portion of a frequency listing for a file including words that a sample of 317 UIC students would enter into a voice response telephone system for student information.

Normally, WORDLINK strips out punctuation and numerical digits from word strings. An option can stop the window at the end of a sentence. Then, it restarts the word-pair counting at the beginning of the next sentence. If desired, specific words can be dropped, for example, common connectors such as prepositions, conjunctions, pronouns, and verbs of being. WORDLINK, however, enables the analyst to choose any set of drop words. A file of them may be input to the program, or none will be dropped.

Table 1. Sample Open-Ended Answers to the Question: "What is your favorite radio station and why?"

@@001 WNUA 95.5 45

WNUA plays a lot of mellow Jazz. On Sunday nights they play a lot of mellow strange tunes. It is a good station to relax, study or think while listening to.

@@002 WVON 1450 25

WVON plays mostly black music rhythm and blues, soul, funk and some rock. I like it because it plays music you can dance to and it features black artists.

@@003 WLNR 106.3 65

They play old and new RandB music, and contemporary jazz. They have all women radio personalities and I believe there are a lot of women in charge of programming. The station is black-owned. The DJ's are not so outrageous that they talk or have gimmicks all through broadcast. They play 50 (or more) minutes of music each hour. I think mostly young women and the older set of adults listen to WLNR. My brother refers to it as a "girls' station" because they play a lot of love songs.

@@004 WGCI 107.5 67

Top 40 format that plays mostly "urban contemporary" music. In other words they play black music like rap and a lot of dance music. The afternoon DJ Tom Joyner is the funniest personality on radio.

@@005 WLUP 1000 10

The station started as an original, seemingly unique idea in radio a few years ago when it began. The goal was not to flood airwaves with another top 40 music format. Instead they found talented radio personalities, mainly humorous ones, that could cater to the 18-52 year old demographic. The FM Loop already had Johnathan Brandmeier, a talented funny, original and charismatic performer. The station rehired Steve Dahl and Gary Meier who have built a cult radio following since 1979 in Chicago with their controversy and their own style of humor. My favorite asset is Kevin Matthews with a very funny and original program where he incorporates dozens of voice characters. He is the most unique and funny radio personality that I have heard. When they do play music it is what I like to hear. Some classic and obscure songs and new cutting edge music.

Table 2. Sample Word Frequency List

WORD	FREQUENCY	WORD	FREQUENCY
registration	72	housing	8
financialaid	62	news	8
grades	46	specialevents	8
sports	45	graduation	7
events	38	movies	7
jobs	28	parking	7
weather	25	teachers	7
tuition	23	time	7
money	21	travel	7
scholarships	20	advising	6
activities	17	employment	6
sex	16	libraryhours	6
classes	14	transcripts	6
concerts	14	campusactivities	5
music	13	careers	5
parties	13	directions	5
counseling	12	financial	5
courses	12	recreation	5
information	12	security	5
clubs	10	tutoring	5
library	10	women	5
food	9	admissions	4
help	9	aid	4
professors	9	athletics	4
entertainment	8	bars	4
health	8	books	4

3) Do Word Cooccurrence Analysis

The next step in the process is to index word cooccurrences. WORDLINK counts the number of times each word in the text occurs at a certain distance from other words. WORDLINK can be set to treat word pairs as directional or non-directional. In the latter case, an A-B pair is treated the same as a B-A one. In the directional option these pairs are kept separate. Retaining word-order information is useful in constructing optimal messages. This is a sort of statistical "unparser" to put messages together at an aggregate level, after breaking down large volumes of input text into word pairs. Table 3 presents a link list of words linked to a focal word. These were from full texts of stories about the U.S. air attack on Libya that appeared in the Spanish newspaper, El Pais.

Table 3. Link List For a Node: Spanish Newspaper Stories About USA Air Attack on Libya

LINKS OF NODE #	5	(LIBYA	)	
TO NODE #	6	(APOY	)	STR-- 9
TO NODE #	7	(TERRORISMO	)	STR-- 36
TO NODE #	38	(A	)	STR1296
TO NODE #	59	(SE	)	STR-- 36
TO NODE #	82	(SIDO	)	STR-- 9
TO NODE #	89	(HAN	)	STR-- 9
TO NODE #	96	(HA	)	STR-- 9
TO NODE #	112	(Y	)	STR-225
TO NODE #	142	(CAPITAL	)	STR-- 9
TO NODE #	149	(MILITAR	)	STR-- 36
TO NODE #	152	(NORTEAMERICA	)	STR-- 49
TO NODE #	153	(CONTRA	)	STR-841
TO NODE #	160	(TRIPOLI	)	STR-- 9
TO NODE #	191	(ESPANA	)	STR-- 16
TO NODE #	197	(EEUU	)	STR-- 25
TO NODE #	256	(ATAQUE	)	STR-169
TO NODE #	287	(CRISIS	)	STR-- 16
TO NODE #	334	(REPRESALIAS	)	STR-- 9
TO NODE #	388	(ES	)	STR-- 16
TO NODE #	429	(CONDENA	)	STR-- 16
TO NODE #	432	(ESTAD	)	STR-- 36
TO NODE #	433	(UNIDOS	)	STR-- 9
TO NODE #	456	(DIPLOMA	)	STR-- 25
TO NODE #	513	(BOMB	)	STR-- 16
TO NODE #	575	(CREO	)	STR-- 9
TO NODE #	666	(SANCIONES	)	STR-- 25
TO NODE #	776	(AMENAZS	)	STR-- 25
TO NODE #	1040	(ATAC	)	STR-- 36
TO NODE #	1706	(REVOLUCION	)	STR-- 9
TO NODE #	1752	(ACUSA	)	STR-- 9

To conceptualize how word-pairing is done, think of a window sliding through the text. It centers on a word, and each pair of words that appears around it within a specified distance is counted. Then the window moves to center on the next word, and the next word, until it has slid all the way through all the text for all the cases. The window is actually a range over which word-pairing occurs throughout a body of text. It refers to the quantity of link steps spanned for word-pair tallying. The upper limit for the quantity of word pairs is (windowsize - 1) x (# of words - 1).

Table 4. Word Pairs and Frequencies: 1990 Flood Stories.

WORD A	WORD B	FREQ
flood	victims	17
state	flood	14
flood	response	12
response	activities	12
for	flood	8
p.m.	update	8
washington	state	8
activities	p.m.	6
flood	damage	5
to	flood	5
flood	insurance	4
flood	recovery	4
aid	to	3
approved	for	3
assistance	approved	3
disaster	unemployment	3
due	to	3
flood	control	3
flood	proofing	3
in	flood	3
ohio	town	3
on	flood	3
recovery	tips	3
unemployment	assistance	3
victims	to	3

Freeman and Barnett (1991) took the same set of data and tested my windowing software against Woelfel's CATPAC, which associated all words within paragraphs. They reported that WORDLINK gave results with higher face validity, interpretability, and applicability. As a result, CATPAC's basic associative algorithm was changed so that it now incorporates the windowing method.

Word-pairs within a window can be treated as having the same strength, regardless of their distance from one another within the window. Or, WORDLINK can weight pair strength based on how close to one another words in a pair are. That is, with text "A B C D", A-B may be weighted at 1.0, A-C at 0.67 and A-D at 0.33.

The outcome of this windowing is a list of word pairs and the total number of times

each pair has occurred within the window across all text. Table 4 shows the most frequent word pairs for a set of news story headlines about floods in the 1990 National Newspaper Index.

The width of the window can vary. In a series of tests, the largest one run was a radius of 20 words on either side of a word, which is a window width of 41. Each level of radius down to one, has been systematically tested. Figure 3 portrays alternative window sizes. The same results held for a radius of 20 down to 3. At radius 2 and radius 1 there were qualitatively different results. This curve was replicated using many different text files. As a result, if the objective is identification of word clusters only, then a radius of 3 is recommended. It gives the same results as higher size radii, but is optimally efficient in terms of computing resources. When setting the radius to 1, word pairs are constituted only by words appearing directly next to one another. With this, when constructing the aggregate message, the gist of the text corpus, there is no ambiguity as to what words should be connected to other words.

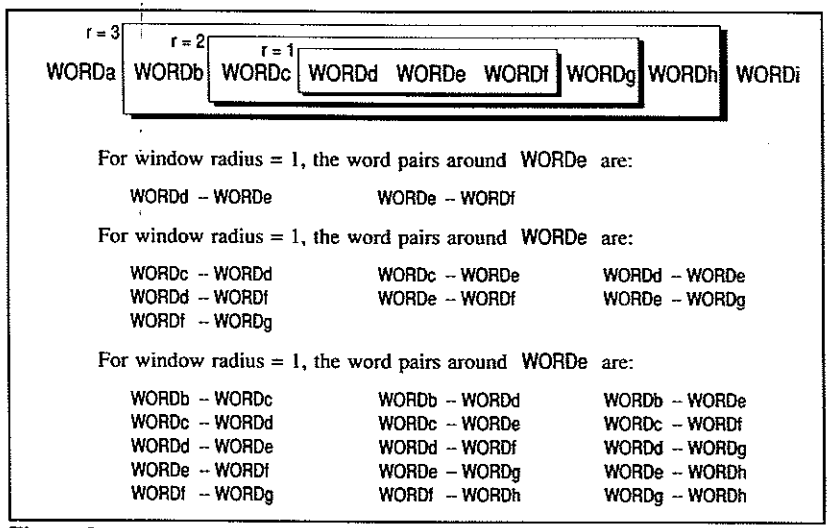


Figure 3

Some readers may question the fact that words are counted more than once with this windowing procedure. They are not, however, counted more than once for listings of word frequencies for individual words. In the pair counting it is a different story. It is true that all words except for the last word are represented more than once. This is not, however, redundant measurement, because the interest is in examining each word's relationship with all the words surrounding it within a given distance. Think of the window concept as a radius pivoting around a word. Take the following text: A B C D E F G H. Let us

examine the relationship between word D and the associated words within a window size of three. The group of words would consist of B C D E F. Given a window size of three D would be counted in the following pairs: B-D, C-D, D-E, and D-F. So, yes there are duplicate, triplicate, and beyond counts of a given word in a text stream. However, any given word-pair is counted only once, unless that pair occurs more than once.

Because WORDLINK has been written to handle text files with up to 1 billion words, sometimes the word-pair output can be so large that it overloads computer system defaults for acceptable file sizes. To get around this problem, a "word grabber" option was implemented. The program can be set to use a file of specified "grab" words. It will grab pairs only around these specific words, using whatever size window is set. For example, if you were interested in scanning corporate annual reports to see how the word "innovative" appeared in context, you could set WORDLINK to turn on its word-pair output generation only when it encountered that particular word. It would then generate output for all word pairs occurring within the window size around "innovative" throughout the text. The current version will output all word pairs appearing within up to 64 words on either side of the grab word.

Table 5a. Group Distance Information, From Open-Ended Responses to Questions about Voice Mail.

distance matrix for group 1, using non-directed links
(the entry is the shortest path from row to column)

ID#	3	13	60	87	91	100	108	121	123	132	137	152	222	223	234
3	0	1	2	2	1	3	1	2	1	2	2	2	2	2	1
13	1	0	2	2	1	3	1	2	2	2	2	2	2	2	1
60	2	2	0	3	3	3	1	2	2	3	2	1	2	2	2
87	2	2	3	0	2	3	2	3	1	2	2	3	3	2	1
91	1	1	3	2	0	3	2	3	2	2	2	3	3	2	1
108	1	1	1	2	2	2	0	1	1	2	1	1	1	1	1
121	2	2	2	3	3	1	1	0	2	1	2	1	2	1	2
123	1	2	2	1	2	3	1	2	0	2	1	2	2	2	1
132	2	2	3	2	2	1	2	1	2	0	2	2	3	2	1
137	2	2	2	2	2	3	1	2	1	2	0	2	1	2	1
152	2	2	1	3	3	2	1	1	2	2	2	0	2	2	2
222	2	2	2	3	3	3	1	2	2	3	1	2	0	1	2
223	2	2	2	2	2	2	1	1	2	2	2	2	1	0	1
234	1	1	2	1	1	2	1	2	1	1	1	2	2	1	0

GROUP COL MEAN= 1.867 STANDARD DEVIATION = 0.3077

4) Do Word-Network Analysis

The lists of word pairs and their frequencies of cooccurrence are then fed to network analysis procedures that can identify a reduced structure. Usually, NEGOPY (Richards & Rice, 1981) is used because it can handle many nodes, 6000 in the mainframe version, and 3000 on a PC. Other network analysis procedures are limited to an order of magnitude fewer nodes, in the range of 100 to 300 maximum. Correspondence analysis has also been useful (Danowski, 1989; Barnett, this volume), as has cluster analysis. The problem, however, is that most implementations limit the number of variables to no more than 500, not 5000 or more, as in NEGOPY.

Table 5b. Analysis of Group Distance Information, From Open-Ended Responses to Questions about Voice Mail.

NODE #	ROW MEAN	STANDARD DISTANCE	WORD
108	1.286	-1.888	i
234	1.357	-1.656	to
223	1.714	-0.495	the
123	1.714	-0.495	leave
3	1.714	-0.495	able
137	1.786	-0.263	messages
121	1.786	-0.263	it
13	1.786	-0.263	am
152	1.929	0.201	not
132	1.929	0.201	me
222	2.071	0.666	that
91	2.143	0.898	get
60	2.143	0.898	do
87	2.214	1.130	for
100	2.429	1.826	has

5) Identify Minimum Spanning Tree

The next step is to find the structure that provides the minimum number of link steps connecting all nodes in the network, the minimum spanning tree. There are different programs in the operations research networks literature that can do this. Nevertheless, NEGOPY is convenient. It identifies minimum path distances among nodes within a group. Table 5a,b provides an example of a NEGOPY distance matrix for a network generated from an analysis of people's descriptions of their use of voice mail (Danowski & Rice, 1989).

6) Mapping Coordinates of Nodes in Euclidean Space

Before implementing a traveling salesman algorithm to identify optimal word strings, spatial coordinates representing the dispersion of the nodes in the network are needed. One cannot run word-pair data directly into a distance model without severe violation of assumptions, and highly questionable solutions. This is because word-pairs frequently violate the triangular equality assumptions of Euclidean geometry (Woelfel & Barnett, 1982). Word A may be linked to B at a high rate; B may be linked to C at a high rate; yet, A and C may never appear within the window together. Nevertheless, after the minimum path distance among all nodes is computed by a network analysis program, these distances are perfectly Euclidean. They can be effectively input to a metric multidimensional scaling program such as GALILEO (Woelfel & Fink, 1980). A graph of word coordinates identified by it appear in Figure 4. These data are from U.S. machine tool buyers' perceptions of Spanish equipment.

7) Traveling Salesman Optimal Solution

The traveling salesman algorithm finds an optimal sequence of nodes such that by visiting each in this order, the network is most efficiently traveled. For reasons discussed earlier, this has theoretical appeal for the model of communication vehicles and optimal message creation.

A useful traveling salesman algorithm is Nagel's (1990) implementation of the classic Bell Labs optimal solution. It requires that each node be identified with a set of spatial coordinates. This is what necessitates the previous step of finding the Euclidean distances among all words based on the minimum path distances among them.

The traveling salesman algorithm was applied to a problem with a sample (n = 500) of automobile dealership personnel (Danowski, 1989). The dealerships were stratified into four cells based on being in either the top or bottom quartiles of the J.D. Powers "Customer Satisfaction Index (CSI)," and by dealer sales volume. Respondents, representative of main dealership roles, including sales, service, parts, and management, were asked the following open-ended question (plus three probes) in a telephone interview survey: "When you think of 'customer satisfaction,' what comes to mind?" Verbatim responses were cap-tured via a computer-assisted telephone interviewing laboratory. Human inter-viewers read questions that appeared on a computer screen, and they typed the word-for-word answers that respondents gave.

The dealerships in the upper half of both the CSI and the sales volume distributions were segmented out for word-network analysis and optimal message creation. Using the procedures described above, the traveling salesman algorithm was applied. It ran on the minimum path distances among the primary group of words resulting from the word-network analysis of the verbatim responses in this cell of the design (n=125).

The traveling salesman algorithm identified the following string of words in this order: "Customer satisfaction: good service on the new car done right the time first." When the order of the last two words is reversed, the message has linguistic validity. It

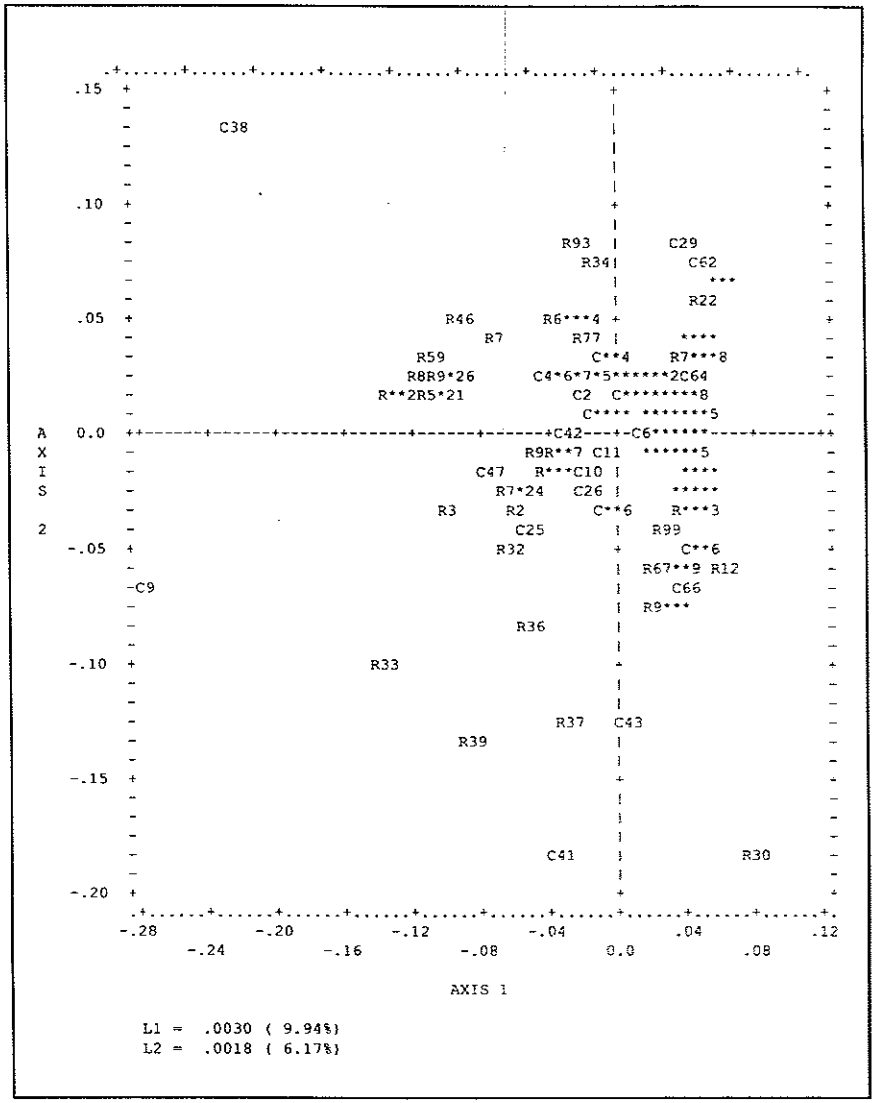


Figure 4. Word coordinates produced by GALILEO for data from U.S. machine tool buyers' perceptions of Spanish equipment

would be understandable to car dealer personnel who might see the message on posters in the dealership, printed on payroll checks, in employee publications, in videos, and in other media.

8) Compute Net Value of Messages

Once a body of text has been network analyzed, one can add side data that indexes value. Value data can be diversely conceptualized depending on the problem at hand:

a) collective-subjective, as exemplified by stock price-to-earnings ratios for organizations;

b) collective-behavioral, such as market share for products/services; or

c) individual-subjective ratings of value, for example, the Customer Satisfaction Index (CSI) for automobiles and dealerships.

After a value metric has been selected, for each case a record is created including frequency variables for each word-pair identified from the initial window scanning step, and for the value index. Then, correlations are run to yield a value weight for each word-pair. With the semantic network now value-weighted, messages can be tested for net value by extracting and summing their constituent word-pair weights from the aggregate network.

Any message can then be evaluated for its net value within that particular semantic domain. It is the sum of the n-step weighted links activated by the words in the message. There is a decay function across successive n-steps. (This decay function can be empirically estimated in an experimental design.) At each step, the value weights are multiplied by the word-pair cooccurrence values to yield a link value/frequency weight. These are summed across all the links for a word within a step-level and multiplied by the step decay value. Values across levels for a word are summed. The result is a subaggregate value weight associated with the particular word in the test message. This process is repeated for each word in the message. The sum is computed for the subaggregate weights for all the words to arrive at the net aggregate value weight for the message.

The formula for net value weight of a message is:

$$NVW = \sum V_i * V_j ((F_{ij} * T_{ij})(\frac{1}{m^E}))$$

where: V_i = value weight for word i; V_j = value weight for word j; F_{ij} = frequency of cooccurrence of ij; T_{ij} = number of two-step links for the ij pair; M = minimum path distance from trigger word (n-step link value); E = activation decay exponent.

This definition of message value, in semantic network terms, takes into account the micro-level relationships among message elements in terms of their associations and

relative value. Such an approach allows the analyst to establish a value reference network for a semantic domain against which any messages can be checked. So, the analyst can compare alternative ones using a common standard. This makes it possible to determine whether qualitatively different messages are significantly different in value.

9) Evaluate New Text Against the Value-Weighted Reference Network

There are two main ways to approach preparing the reference network. One way is when the target population is clearly known in advance. The other is when the population is not so clearly identifiable but the criterion variables of interest are. For example, an organization may already tag the group of high performers by other procedures, such as quarterly or annual ratings. If a suitable identification of the reference group does not already exist, then the analyst could use a different approach: sample from the whole range of units tested; for example, salespersons. In addition to their open-end question answers, the analyst obtains reliable and valid data on the criterion variable of interest. In this case, the variable would be some measure(s) of sales performance, such as total volume of sales, or supervisor ratings. First, the analyst identifies the semantic network for the whole sample. Then, he or she value weights it on the criterion variable. Next, the correlation coefficient for each word pair with the value criterion variable serves as a weight for that pair in the network.

Then, the analyst gathers open-end data from test subjects using the same reference question(s). The test subjects, newly hired prospects, for example, are asked the exact same question as was asked in the reference sample. Their answers are recorded verbatim.

The fit test comes next. Each individual's answers are processed to measure the distance of each word from the center, and the reference weights are assigned. The values are summed for each word encoded by the test subject. These total scores represent the closeness of each subject to the reference network.

Decisions can then be made about whatever social unit is being evaluated -- a person, an organization, or whatever. There are different decisions possible, depending on the overall goals. One goal may be to place individuals into cohesive units. Another may be to identify people to serve as liaisons between groups. Or, a goal may be to introduce change into the organization, in which case one would select the more peripheral, rather than the semantically central individuals.

10) Test Predictions

The network analysis of open-end text provides a way to do precise time-series analysis with qualitative data. The actual change in word-networks can be specifically tracked. This testing is useful for two kinds of applications. One is optimal message creation and dissemination. The other is selection of social units according to fit with a reference word-network. By re-administering the same open-ended questions or gathering texttial samples from these units, we can see how well the predictions were borne out.

11) Update Reference Population, Sample, and Reference Network

Over time, it would be good to revisit the reference population, redefine it as necessary, sample it again, and update the reference network. This updating can be part of the long-term, over-time research on word networks. The reference network can be revised, based on predictive validity data, or on externally specified changes in the reference population and sample.

V. FUTURE DEVELOPMENTS

At the theoretical level, one area of development is to take diffusion theory and recast it in terms of language variables. In its first phase of development in the 1960s (Rogers, 1965), diffusion theory broke the social system into different types of actors. They were defined in terms of their position at various stages of the continuous time curve for adoption of the innovation. In the second phase, diffusion theory viewed individuals over time in more network-oriented terms (Danowski, 1976; Rogers, et al., 1976; Rogers & Kincaid, 1980; Rogers, 1983; Danowski, 1986; Rice, Grant, Schmitz, & Torobin, 1990). The theory conceptualized the links between variations in individuals' interpersonal networks and their participation in adopting innovations.

A new, third phase of diffusion theory advancement is possible. This would add the language layer to diffusion. It would treat the propelling forces of word associations in driving social vehicles. It would account for the travel of communication vehicles over time through the social system, in terms of the semantic associations to them.

Evidence exists that people with different interpersonal network structures have different media use patterns, and different psychological orientations. These patterns are consistent with inferences we can draw from diffusion theory (Danowski, 1986). Furthermore, people with different interpersonal network structures use language differently (Danowski, 1988). These word differences are also consistent with diffusion theory. So, adding message variables would help center the diffusion process. To date, words in messages have been examined only grossly, in terms of taboo words that inhibit diffusion. Nevertheless, a natural language approach to full-text messages about innovations is possible with the methods explicated in this chapter. Both the language of the campaign messages, and that of the members of the social system about the innovation, can be sensitively quantified over time. A theory of word waves and social change could be the result: a word-network spin theory.

Observing the network of word-nodes and links over time reveals waves of word activations. At times of least activation, a vehicle may not pass near a set of word nodes. Later, it may begin to move through that region, and to accelerate in how often it does so. This activation looks like the classic s-shaped diffusion curve. Then, after hitting the peak, the deactivation side looks like the reverse. Taking both sides together, activation and deactivation, constitutes a normal distribution of activation over time. Furthermore,

stringing successive S-curves and reverse S-curves together over time gives us a word activation wave. Wave mechanics metrics can index word wave phases, periods, and frequencies. So, taking word trips over time, communication vehicles cut word waves. In order to test message effects hypotheses, the symbolic analyst may want to position the vehicle to catch a wave created by the coherent actions of other vehicles, and to measure the effects on criterion variables.

There are also some methodological fronts to develop. Morphological parsers of text, and parts-of-speech taggers can be further explored. Morphological parsers reduce all the lexical variations of a word down to its basic root. For example, 'uttering, utterance, utterances, utterly' would all be parsed to "utter." Parts-of-speech parsers assign words to their grammatical function categories, such as noun, verb, or adverb. To date, however, this researcher has rejected either form of parsing. Rather than reduce natural language to some simpler form, every word is used as it appears (after spelling corrections). The value of using every word in its original raw form became clear during the study about automobile dealerships and customer satisfaction. One of the first analyses done was a simple test for differences in word frequencies, comparing the high to the low CSI-rated dealerships. It was surprising to find that the plural word, 'customers' was significantly more frequent for the highly-rated dealerships. In contrast, the low-rated dealerships used the singular form.

Finding this subtle difference led us to look at the associations with both the singular and plural forms in the word networks. It was revealed that the plural form linked to words suggesting a model for how customers as a group fit into the business plan of the dealership. The highly rated dealerships appear to attempt to treat every customer the same way, and a good way at that. In contrast, the singular form was related to focusing on an individual customer's problems, questions, needs, complaints, etc. In these dealerships, personnel adapt to each individual customer and tailor their communication behaviors to them. Such an individual customer-specific focus was associated with low ratings by customers. If we had reduced all plural forms of words to their singular roots, we never would have discovered this counter-intuitive finding.

Another unexpected empirical experience bears on the issue of dropping words. Forbes (1991) was analyzing news stories that appeared in Third World, national USA, and local USA sources. Because the pilot test had a small number of news stories, no drop list was used, so that maximum insight could be gained from the pilot test analysis. The results revealed the unexpected finding that Third World news stories had considerably lower use of the words "to" and "of." At first this seemed like a trivial finding. Upon further reflection, however, the dictionary was examined to see what it had to say about these words. It was thought that perhaps they were so commonplace that something interesting might be inadvertently overlooked. The dictionary definitions of both words made it clear that they were important to specifying relationships among things. Then, this lead to the hypothesis that Third World stories are covered qualitatively differently than others, with perhaps a more simplified treatment, presenting less information about complex relationships than is the case for more local stories. This may

suggest a sort of bias different from that proposed by advocates of the "New World Information Order" -- albeit one for which credible scientific evidence is difficult to obtain. Here, the linguistic results may reveal a subtle sort of bias that presents a simpler view of the Third World than of other worlds. The manifest content may be less of a problem than the manner in which it is presented.

Nevertheless, some recent developments are opening up new vistas for us. These are based on our work with supercomputers for word-network analysis. On a supercomputer, our WORDLINK software can handle texts containing up to 1 billion words, each up to 32 characters in length. The low end is a PC platform implementation with 2 meg of RAM that can process texts with up to 1 million words. (See Table 6 for more details.)

Table 6. Limits on Current WORDLINK Software

Platform	Memory	Words	Output pages
Super Workstation (RISC 6000-550)	128m RAM	1,000,000,000	200,000
Mainframe machine	32m RAM	25,000,000	50,000
Mainframe machine	16m RAM	12,500,000	25,000
Mainframe machine	8m RAM	6,250,000	12,500
PC	4m RAM	2,500,000	5,000
PC	2m RAM	1,000,000	2,000

The supercomputer exercise has been fruitful, because it has pushed the envelope of our analysis techniques out on other fronts. We have also updated our software as we converted it from SPITBOL to C, required because SPITBOL is not available on super workstations, and we added new features and options.

Nevertheless, the supercomputer project has created some interesting problems. We can now do the word-pair analysis using WORDLINK with files that generate an order of magnitude more word-pair data than we can fit to our currently largest-capacity structural analysis program, NEGOPY-EQN for mainframes (Richards, 1985). The highest capacity NEGOPY version currently available cannot handle more than 80,000 different word-pair links. So, as next steps in our word-network content analysis research program, we are looking to morphological and parts-of-speech parsing to reduce the newly created overcapacity data to under 6,000 unique words and 80,000 links. As well, a high capacity cluster analysis program is currently under development.

Conclusion

This chapter focused on using network analysis to analyze message texts. It reviewed some of the limitations of prior related work. Then, it outlined the development of a set of procedures for word-network analysis. Next, it turned to previewing future developments. As a final note, Table 7 outlines some benefits of word-network analysis.

Table 7. Benefits of Word-network Analysis of Self-report Verbatims.

- Represents the semantic content of messages in the actual, natural language in which they were originally expressed, resulting in greater external validity;
- Reduces translation error in moving from what is said to the representation, resulting in greater internal validity;
- Preserves relationships among concepts in the coded representa-tions, resulting in a qualitative analysis that is quantified, enabling the integration of the two kinds of analysis;
- Retains more information compared to traditional nominal and manual coding, which often reduces rich qualitative information to five to nine nominal categories;
- Removes biases of human coders;
- Minimizes the number of coding transformations between the original utterances, and the final representation that is statistically analyzed, thus reducing analytical error;
- Processes large volumes of open-ended or message text data well beyond the capacity of human coding; and
- Enables precise quantitative comparisons of qualitative information across different groups of respondents.

References

Alexander, M. & Danowski, J. (1990). Analysis of an ancient network. *Social Networks*, 12, 313-35.

Danowski, J. (1991a, April). Radio semantics: Audience word-networks as predictors of market share changes. Paper presented to the National Association of Broadcasters, Las Vegas.

Danowski, J. (1991b, February). Collective semantic network structure and content as predictors of market share changes. Paper presented to the International Network for Social Network Analysis Conference, Tampa, Fla.

Danowski, J. (1990). Organizational media theory. *Communication Yearbook*, 14, 187-210.

Danowski, J. (1988). Organizational infographics and automated auditing: Using computers to unobtrusively gather and analyze communication. In G. Goldhaber and G. Barnett (eds.) *Handbook of organizational communication* (pp. 335-384). Norwood, NJ: Ablex.

Danowski, J. (1982). A network-based content analysis methodology for computer-mediated communication: An illustration with a computer bulletin board," *Communication Yearbook*, 6, 904-925.

Danowski, J. (1980a). Formative computer conference message evaluation methodology. *Proceedings of the Conference on Evaluating Computer Conferencing*, R. Hiltz & Turoff, M. (eds.) Cranford, NJ.

Danowski, J. (1980b). Message Content Cooccurrence in Computer Communication Networks. Invited presentation to the Workshop on Metric Multidimensional Scaling, International Communication Association, Acapulco, Mexico.

Danowski, J. & Martin, T.H. (1979). Evaluating the health of information science: *Research community and user contexts*. Final report to the Division of Information Science of the National Science Foundation, no. IST78-21130.

Danowski, J. & Rice, R. (1989, May). Correspondences between users' semantic networks and computer-monitored data on users of voice mail for messaing versus answering. Paper presented to the International Communication Association.

Forbes, S. (1991). *Word networks across first, second, and third world newspaper stories*. Unpublished masters thesis. University of Illinois at Chicago.

Freeman, C. & Barnett, G. (1991, May). An alternative approach to using interpretive theory to examine corporate messages and organizational culture. Paper presented to the International Communication Association, Chicago.

Hall, S. (1988). *The hard road to renewal : Thatcherism and the crisis of the left*. London; New York: Verso.

Hockey, S., Burnard, L. & Sperberg-McQueen, M. (1991, April). Status report: Text encoding initiative. *Proceedings of the ACH/ALLC conference*, Phoenix, 205-208.

Krippendorff, K. (1980). *Content analysis : An introduction to its methodology*. Beverly Hills, Sage Publications.

Lewin, K. (1931). The conflict between Aristotelian and Galilean modes of thought in contemporary psychology. *Journal of General Psychology*, 5, 141-177.

Nagel, R. (1990) Optimal traveling saleman algorithm. Computer program.

Richards, W., Jr. (1986) *The NEGOPY Network Analysis Program*. Monograph, Simon Fraser University, Burnaby, B.C., Canada.

Richards, W. Jr. & Rice, R.E. (1981). NEGOPY network analysis program. *Social Networks*, 3(3), 215-223.

Rogers, E.M. (1965). *Diffusion of innovations*. New York: Free Press.

Rogers, E. M. (1983). *Diffusion of innovations*. 3rd ed. New York: Free Press.

Rogers, E.M. & Kincaid, D.L. (1981). *Communication networks: toward a new paradigm for research*. New York: Free Press.

Rice, R.E., Grant, A., Schmitz, J. & Torobin, J. (1990). Individual and network influences on the adoption and perceived outcomes of electronic messaging. *Social Networks*, 12(1), 27-55.

Stone, P.J., Dunphy, D.C., Smith, M.S. & Oglivie, D.M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, Mass: M.I.T. Press.

Weber, R.P. (1990). *Basic Content Analysis*. Newbury Park, Calif.: Sage Publications.

Wilks, Y. (1989). *Theoretical issues in natural language processing*. Hillsdale, N.J.: L. Erlbaum.

Woelfel, J. & Barnett, G. (1982). Multidimensional scaling in Riemann space. *Quality and Quantity*, 16, 469-491.

Woelfel, J. & Fink, E. (1980). *The Galileo system: A theory of social measurement and its application*. New York: Academic Press.

Zull, C., Weber, R. & Mohler, P. (1989). *Computer-aided text classification for the social sciences: The General Inquirer III*. Monograph, Zuma, Mannheim, Federal Republic of Germany.