Quality & Quantity **34:** 259–274, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

A Conceptual Framework for Quantitative Text Analysis

On Joining Probabilities and Substantive Inferences about Texts

CARL W. ROBERTS
Department of Sociology or Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

Abstract. Quantitative text analysis refers to the application of one or more methods for drawing statistical inferences from text populations. After briefly distinguishing quantitative text analysis from linguistics, computational linguistics, and qualitative text analysis, issues raised during the 1955 Allerton House Conference are used as a vehicle for characterizing classical text analysis as an instrumental-thematic method. Quantitative text analysis methods are then depicted according to a 2 × 3 conceptual framework in which texts are interpreted either instrumentally (according to the researcher's conceptual framework) or representationally (according to the texts' sources' perspectives), as well as in which variables are thematic (counts of word/phrase occurrences), semantic (themes within a semantic grammar), or network-related (theme- or relation-positions within a conceptual network). Common methodological errors associated with each method are discussed. The paper concludes with a delineation of the universe of substantive answers that quantitative text analysis is able to provide to social science researchers.

Key words: content analysis, text analysis, semantic grammar, network, instrumental versus representational, quantitative methods.

1. Introduction

Painted with broad strokes, formal analyses of linguistic data are pursued from three academic orientations: linguistics, computer science, and the social sciences. Linguists' interests are primarily in describing text structure: as surface forms produced by an innate human capacity to generate linguistic expressions (Chomsky, 1965), as sequences of functional forms expressed by goal-oriented humans according to discrete narrative grammars (Griemas, 1984 [1966]; Halliday, 1994), as patterns of symbols uttered by fallible native speakers in ways (in)consistent with some prescribed standard (Honey, 1983), etc. With recent developments in computer technology, linguists have begun to evaluate their theories by developing corresponding text-parsing software (cf. Rosner and Johnson, 1992). This work forms the academic branch of computational linguistics, in addition to which a more applied, commercial branch has developed with the objective of quickly "understanding" user input and yielding as output the user-expected outcome (Grishman, 1986; McEnery, 1992). Finally, social scientists conduct formal analyses of written and spoken texts to reveal mechanisms according to which words influence and are influenced by human behavior.

This paper is a review of *quantitative text analysis methods* – one of two general classes of methods currently in use for the social scientific analysis of texts. To be quantitative, a text analysis must both address a social scientific question of a well-defined text population, and provide an answer to the question having a known probability of inaccurately reflecting aspects of the text-population.[1] Although there is presently an astounding number of recent books on qualitative text analysis (e.g., Fielding and Lee, 1991; Riessman, 1993; Silverman, 1993; Denzin and Lincoln, 1994; Feldman, 1994; Krueger, 1994; Marshall and Rossman, 1995; Miles and Huberman, 1994; Wolcott, 1994; Kelle, 1995; Weitzman and Miles, 1995), virtually all discussions of quantitative text analysis methods are written as if there had been no innovations in these methods since the 1960s (cf. Altheide, 1996; Lee, 1999; but, as an exception, Roberts, 1997a). In the following section, I use the mid-1955 Allerton House Conference debates on contingency analysis to introduce the classical approach to text analysis that then and into the 1980s was the most predominant text analysis methodology in the United States. The approach's instrumental orientation is then differentiated from an increasingly utilized representational text analysis orientation. To complete my 2 × 3 classification of quantitative text analysis methods, I then distinguish classical thematic text analysis from more recent semantic and network text analysis methods.

This paper's purpose is to impose some long-needed structure on a wide spectrum of text analysis methodologies that heretofore have been accessible only among a smattering of methodology journals. Its emphasis is on application. That is, it is intended to aid researchers in answering the question, "Which quantitative text analysis method best affords answers to what research question?"

2. Classical Text Analysis

Quantitative text analysis has a long tradition in the works of Lasswell, Berelson, George, Osgood, Pool, Stone, Holsti, Krippendorf, Weber, and many others. During a work conference held in 1955 at the University of Illinois-Monticello's Allerton House many of these text analysis pioneers gathered to develop solutions to the methodological problems of the day. *Trends in Content Analysis* (Pool, 1959) is the scholarly legacy of this conference.

The most influential, if not the largest faction among the conference's participants was a group of Harvard researchers who made extensive use of what they called "contingency analysis". The first step in a contingency analysis involves counting occurrences of content categories within sampled blocks of text. This produces a data matrix like that in Table 1, with distinct content categories (or themes) heading the columns, unique text blocks heading the rows, and counts of occurrences (of theme within block) in the cells. The analysis proceeds by computing a matrix of associations between pairs of themes. Finally, the researcher

Table I. A data matrix for a thematic text analysis

ID-number	Theme 1	Theme 2	Theme 3
1	2	0	0
2	0	0	1
3	1	3	1
4	0	2	1
5	0	0	0
.	.	.	.
.	.	.	.
.	.	.	.

develops (usually post hoc) explanations of why some themes co-occurred and why others were disassociated (i.e., negatively associated).

During the Allerton House conference Alexander George (1959: 17f), whose work on World War II propaganda did not use contingency analysis, pointed out that such "fishing expeditions" (sic) are not sensitive enough to detect the instrumental use to which communication is put by the speaker. Co-occurrences do not reveal how themes are used in the same blocks of text, let alone why the speaker used them that way. In response, exponents of contingency analysis acknowledged that their technique is not able to detect changes in communication strategies (i.e., it cannot legitimately be used to investigate instrumental communication), but that it can be used to trace patterns in representational communication (i.e., communication that "means what it says on its face"). Thus classification of representational communication into content categories is done based on the assumption that "what an author says is what he means" (Pool, 1959: 4).

Amidst this curious exercise in turf delineation, Charles Osgood (1959) described and illustrated his evaluative assertion analysis – precursor of contemporary network text analysis – in building a defense for analyzing the representational content of communication. Here, in a casual but extremely insightful remark, Osgood (1959: 75) notes, "As a matter of fact, we may define a method of content analysis as allowing for 'instrumental' analysis if it taps message evidence that is beyond the voluntary control of the source and hence yields valid inferences despite the strategies of the source". And later regarding contingency analysis, "The final stage, in which the analyst interprets the contingency structure is entirely subjective, of course" (p. 76). With these remarks, Osgood attached an entirely different meaning to the term, instrumental. No longer did it allude to the strategy behind a source's communication, but to the researcher's *interpretive strategy* in analyzing the communication. In this usage, words are not the source's strategic instruments, but are symptomatic instruments from which the researcher can diagnose possibly unconscious or unacknowledged characteristics of the source.

3. Representational Versus Instrumental Interpretation[2]

Shapiro (1997) has recently extended Osgood's sense of "instrumental analysis" by differentiating it from "representational analysis" in which the researcher attempts "to classify, tag, or retrieve the intended meanings of the authors" (p. 228). At issue in this distinction is no longer whether or not the source's intended meaning is part of a tacit strategy, but whether it is the source's or the researcher's perspective that is used to interpret the texts under analysis. When a researcher understands texts representationally, they are used to identify their sources' intended meanings. When a researcher understands texts instrumentally, they are interpreted in terms of the researcher's theory. Thus Namenwirth exemplifies the latter approach in his argument that the sources of his texts "are unfamiliar with many fundamental properties of their own culture and prove unable to specify its structural rules. . . . (T)o recover culture's properties and rules, we cannot ask culture's participants to answer these questions. Instead, we must rely on outsiders as investigators and use their methods, however unreliable these may prove to be" (Namenwirth and Weber, 1987: 237). Accordingly, instrumental text analysis methods are used to identify individual and societal characteristics about which society members may be unaware; representational methods are used to characterize texts in ways that their sources intended them to be understood.

Osgood and Shapiro's recasting of the representational/instrumental distinction helps direct attention to the import that researchers' analytic frameworks have for their findings. For example, consider a small subpopulation of male novelists who only write books with elderly female heroines. In seeking to understand their choices of leading characters, the researcher with a representational orientation might scour the novelists' prefaces, possibly discovering references there to a need for elderly female heroines as a corrective to an overuse of young male heroes in contemporary literature. However, according to a researcher with an instrumental, Freudian orientation, the choice of matronly heroines would likely be symptomatic of Oedipus complexes among the novelists. Note that in the former case the novelists' own meanings are assigned to their writings, whereas in the latter case the researcher's (Freudian) perspective is applied. Given Osgood's and Shapiro's recasting of the representational vs instrumental distinction, classical text analysis (in which content categories based on the researcher's perspective are used in generating matrices of theme-counts per block of text) can now be classified as "instrumental thematic text analysis" – one of six text analysis methods to be introduced in this paper.[3]

4. Data Matrices Produced in Quantitative Text Analysis

Independent of whether it is representational or instrumental, a quantitative text analysis always involves the production of a data matrix. That is, quantitative text analysis requires that words be mapped into a two-dimensional matrix representation suitable for statistical analysis. This fact greatly simplifies the task of delineating the domain of possible questions that quantitative text analyses are able to address. In particular, it allows the sought-after domain to be defined according to the various ways in which the columns (or variables) and rows (or units of analysis) of this data matrix can be defined. As this domain gains clarity, a "theoretical map" emerges on which text analysts can locate both their substantive question, and the text analysis technique(s) to which it corresponds.

4.1. VARIABLES

During the past few decades semantic and network text analysis methods have added to classical text analysis's word/phrase counts, other types of variables for the statistical analysis of linguistic data. Whereas in a thematic text analysis (of which classical text analysis is an instance) one examines occurrences of themes, in a semantic text analysis the examination is of sentences (or clauses) in which themes are interrelated. Moreover, in a network text analysis the examination is of themes' and/or sentences' locations within networks of interrelated themes. The three are not mutually exclusive, but may be combined in the same analysis.

4.1.1. *Thematic Text Analysis*

The type of data matrix generated in a thematic text analysis has already been described in conjunction with Table 1. In brief, it is a matrix having one row for each randomly sampled block of text, and one column for each theme (or concept) that may occur in these text blocks. Cells in the data matrix indicate the number of occurrences of a particular theme within a specific block of text. These data are occasionally embellished with secondary variables that measure the source's positive or negative sentiment regarding each theme (cf. Pool, 1955; Holsti, 1969; Smith, 1992). As computer power has grown, and key-word-in-context searches have become easier, researchers have developed text analysis software with which ad hoc dictionaries (i.e., sets of content categories) can be constructed interactively (Popping, 1997: 214–16). When the themes in these dictionaries are constructed to reflect the meanings intended by the texts' sources, consequent analyses are "representational thematic text analyses". When dictionaries' themes are constructed to reflect the researcher's perspective for interpreting the texts, consequent analyses are "instrumental thematic text analyses" (a.k.a. classical content analyses). In all cases it is essentially a matrix of word-counts that forms the basis for a thematic analysis of texts.

Analyses of word-counts yield inferences about the predominance of themes in texts. For example, Namenwirth and Weber's (1987) cultural indicators research reports shifts in the prevalence of various political and economic themes over time. Yet note that if, for example, certain types of "political protest" are found to occur in texts in which one also finds mentions of "economic inflation", one is unable (on the basis of a data matrix of word-counts) to determine if the protests are *mentioned in the texts as* being the cause or the effect of inflation. This is because information

Table II. A data matrix for a semantic text analysis

ID-number	Subject	Action	Object
1	11	35	64
2	9	22	89
3	14	35	72
4	11	36	72
5	17	30	55
.	.	.	.
.	.	.	.
.	.	.	.

on semantic relations among themes such as "political protest" and "economic inflation" is not afforded by aggregated word-count data.

4.1.2. *Semantic Text Analysis*

Relations among themes are encoded in a semantic text analysis, however. In a semantic text analysis the researcher begins by constructing a template (a.k.a. a semantic grammar). Themes from sampled texts are then mapped as syntactic components within this template. For example, Markoff, Shapiro, and Weitman (1974) developed a two-place semantic grammar for "grievances", that contained one syntactic component for the object of the grievance (i.e., what is being grieved about) and another for the action that should be taken toward this grievance. The reader is referred to Franzosi (1990) and Roberts (1997c) for more detailed descriptions of the application of semantic grammars in text analysis.

Table 2 illustrates a data matrix that might have been generated in a semantic text analysis. Note that the cells in the data matrix do not contain indicators of theme occurrences, but contain discrete codes for the themes themselves. The column in which a specific theme's code appears indicates the theme's syntactic role within the researcher's semantic grammar. In generating Table 2, blocks of text were encoded as sequences of subject-action-object triplets. Inferences from such a data matrix might be made within randomly sampled newspaper accounts of labor disputes, by comparing the odds that representatives of management versus of labor initiate collective bargaining. Within transcribed speech from a sample of minutes of prime-time television content, inferences might be drawn regarding the odds that blacks versus whites refer to themselves as targets of aggression. More generally, and in contrast to thematic text analysis, semantic text analyses yield information on how themes are related according to an a priori specified semantic grammar.

Writings on semantic text analysis methods date back Gottschalk's instrumental analyses in the 1950s on psychological states and traits (e.g., Gottschalk and Kaplan, 1958). Gottschalk has since developed highly automated, parser-based text-encoding software for measuring (according to his perspective on states that are reflected in how people relate words) such psychological states as hostility, depression, and hope (Gottschalk and Bechtel, 1989). The Kansas Events Data System, or KEDS, is another special-purpose, parser-based program that can be used either instrumentally (in conjunction with the standard World Events Interaction Survey, or WEIS, coding scheme) or representationally (based on user input content categories gleaned from one's data) in analyzing event data from news reports (Gerner et al., 1994). Most other uses of semantic text analysis are representational, however. For examples see Franzosi's (1989, 1997a) research on labor disputes, Roberts's (1989, 1991) on ideology shifts, and Shapiro and Markoff's (1998; Markoff et al., 1974; Markoff, 1988) work on public opinion in 18th century France.

4.1.3. *Network Text Analysis*

Network text analysis originated with the observation that once one has a series of encoded statements, one can proceed to combine these statements into a network. Moreover, once text blocks are rendered as networks of interrelated themes, variables can be generated to measure the "positions" of themes and theme-relations within the networks. For example, let us imagine that we construct a network of themes in which all linkages indicate causal relations. Assigning the names theme-A and theme-B to any pair of themes in the network, one could develop a measure of "the causal-salience of theme-A on theme-B" as the proportion of all sequences of causal linkages that are ones in which theme-A is the cause and theme-B is the effect. Note from the simple three-theme network in Figure 1 how calculation of "the causal-salience of theme-A on theme-B" draws on more than isolated semantically-linked themes in blocks of text. It incorporates information on all themes and links within network representations of text-blocks.

Thus a data matrix such as that in Table 3 might be generated from a sample of networks that contained variables measuring the causal-salience of each pair of the texts' themes. Network text analysts have developed many other measures of network characteristics. A theme's "conductivity" is one example, referring to the number of linkages that the theme provides between other pairs of themes (Carley, 1997). Another is of theme linkages that are logically implied, but not explicitly stated in each block of text (Kleinnijenhuis et al., 1997). As of this writing nearly all quantitative network text analysis research has been conducted either by Kathleen Carley or by a group of Dutch researchers in Amsterdam (Carley, 1986; van Cuilenburg et al., 1986, 1988; Carley and Palmquist, 1992). All of these researchers do representational network text analysis. Carley (1988) has even developed expert system software to "fill in" networks with sources' background knowledge.[4]

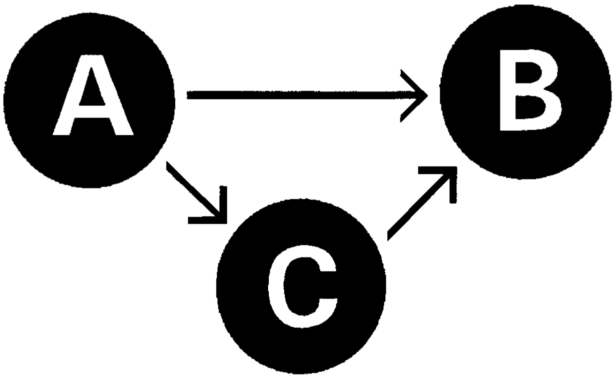

Figure 1. A network of causal relations among themes. (A) There are four sequences of causal linkages in this figure. (1) A → B; (2) A → C; (3) C → B; (4) A → C → B. (B) Note that 0.50 is the proportion of all sequences of causal linkages in which A is the cause and B is the effect. That is, 0.50 is the "causal salience" of theme-A on theme-B.

Table III. A data matrix for a network text analysis

	Causal salience measures					
ID-number	A on B	A on C	C on B	B on A	C on A	B on C
1	0.50	0.25	0.25	0.00	0.00	0.00
2	0.25	0.00	0.50	0.00	0.25	0.00
3	0.00	0.00	0.25	0.25	0.50	0.00
4	0.00	0.00	0.00	0.50	0.25	0.25
5	0.00	0.25	0.00	0.50	0.00	0.25
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

4.2. UNITS OF ANALYSIS

The discussion this far has yielded a 2 × 3 taxonomy of quantitative text analysis methods. These methods are distinguished on the first dimension according to whether the source's or the researcher's perspective is the basis for interpreting texts, and on the second dimension according to whether they use variables that reflect occurrences of themes, themes in semantic roles, or network-positions of themes or theme-relations. Let us now proceed to specify the types of units of analysis that are possible in a text analysis.

In the earliest stages of every quantitative text analysis, the researcher is confronted with a "mountain of words" (a.k.a. a text-population) about which statistical inferences are to be drawn. On the one hand, this text population may be an initially undifferentiated mass. (For example, a sample of minutes of speech on U.S. prime-time television could be drawn at random from the undifferentiated "mountain of words" that were uttered during a year' time.) On the other hand, the text population may consist of clusters of sentences such as newspaper editorials, transcripts of interviews, diary entries, and so on. In either case, a representative sample can only be drawn once the text-population has been divided into distinct text-blocks, each of which is then assigned a unique number and sampled at random.[5]

Yet the text-population should not be mindlessly divided, even if it appears to the researcher as already a collection of discrete text-blocks such as editorials, interviews, or diary entries. This is because the statistical inferences that the researcher may legitimately draw from texts, depend fundamentally on the units into which the text-population is initially divided. Allow me to explain.

Imagine a researcher, who wishes to analyze prime-time television data according to performers' mental models (or conceptual frameworks). In this case, the researcher would begin by dividing the text-population into blocks associated with each *performer*. On the other hand, if the researcher wished to analyze the narratives (or story-lines) depicted on prime-time television programs, the researcher would begin by dividing the text-population into blocks associated with each *program*. Yet performers may appear in more than one program, and programs will involve many performers. As a result, the researcher's inferences will differ, depending on this initial division of the text-population.

In brief, the researcher who wishes to ask a substantive question of a population of texts must not only consider the thematic, semantic, and network variables required to address the question, but also the units of analysis yielded when this population is divided into text-blocks. Thus far I have mentioned two types of units of analysis that might be yielded when a text-population is divided, namely the conceptual framework of the text's source (for example, the television performer's mental model), and the message that the text conveys (for example, the program's story-line). But of course, there are others.

Consider Harold Lasswell's (1948) oft-cited depiction of communications research as the study of "Who says what, in which channel, to whom, and with what effect?" In his article-long answer to this question Lasswell argued that communications' effects can be largely understood as functions of their source, message, channel, and audience. Moreover, these four aspects of communication are the most common *contextual variables* used in analyses of texts and transcripts:

- Characteristics of source (gender, affiliation, biases, etc.)
- Characteristics of message (local vs domestic news, descriptive vs evaluative orientation, etc.)
- Characteristics of channel (radio vs TV news, public vs commercial network, written vs spoken medium, etc.)
- Characteristics of audience (socio-cultural and/or historical setting within which text appeared).

Thus, other than source- and message-identifications, the researcher may also identify text-blocks according to their channel and/or intended audience.

Yet comparisons among texts' sources, messages, channels, and audiences are only possible if each text-block under analysis can be clearly identified according to its type of source, message, channel, and/or audience. The key in selecting a unit of analysis is not to assume that one's population of text is comprised a priori of clearly-distinguishable text-blocks. On the contrary, it is the researcher's responsibility to divide this population into blocks – blocks that can be uniquely identified according to the contextual variables required for addressing the research question at hand.

5. Caveats for the Quantitative Text-Analytically Inclined

In his criticism of contingency analysis, George correctly pointed out that considerable contextual information must be used in interpreting the purposes behind sources' words. Yet apparently lost to many in the Allerton House debates was an awareness that contextual information is necessary to render all communications meaningful. Even if occurrences of thematic categories appear to be valid on their face, these categories' selection must inevitably have been motivated by some theoretical perspective (e.g., the psychological orientation of the generic dictionaries developed for use with the General Inquirer [Stone et al., 1966; Goldhamer, 1969]). Whether the sources' or the researcher's perspectives are used in interpreting texts, these perspectives must be made explicit for the reader to evaluate the validity of conclusions that are made.

Occasionally thematic text analysts will wrongly interpret co-occurrences of themes (i.e., correlations between word-frequencies) as indicative of specific semantic relations among these themes. The warning here is that if one's substantive question is about semantic relations, themes should not be counted, but should be encoded according to a semantic grammar. Indeed it was precisely to overcome contingency analysis's limitation to inferences about theme occurrences that semantic text analysis methods were developed in the first place. Critics have simply lost faith in instrumental thematic text analysts' expertise in divining the "true meanings" of co-occurrence patterns.

For semantic text analysts one typical pitfall stems from the fact that one's units of analysis are often clustered within one's sampling units. For example, in Linguistic Content Analysis (Roberts, 1989, 1997b; Eltinge, 1997) the unit of analysis is always the grammatical clause. Given that prior to sample selection each sampling unit must be assigned a unique number, one's sampling unit will rarely (except for the very smallest of text populations) be the clause. More commonly, one will sample text-blocks (e.g., paragraphs, editorials, etc.) within which clauses are clustered. In such cases statistical analyses should not artificially inflate one's sample size by treating each unit of analysis as if it were a randomly sampled observation. Instead, inferential statistics must be adjusted to take the clustering into account.[6]

A second potential problem in semantic text analysis harks back to the Allerton House Conference, where practitioners of contingency analysis assumed that their thematic categories could be used to capture what words mean on their face, as it were. In parallel fashion, one might be tempted to map thematic relations according to their surface grammatical relations (e.g., as Subject-Verb-Object, or SVO). This approach can work if one's texts are highly descriptive, for example, like lead sentences in wire services' news briefs (Schrodt, 1993; Savaiano and Schrodt, 1997). Yet the intended meanings of much, if not most natural language expressions are inherently ambiguous. The expression, "Joan is a doctor", might be used to state Joan's occupation or to express awe at her accomplishment. "Joe was abandoned", might convey the state of affairs of Joe's being alone or (if passive voice were intended) the process of others leaving him. Encoding these as "S = woman, V = to be, O = medical professional" and "S = man, V = to be, O = without companions" merely passes these sentences' inherent ambiguities on into the encoded data. Analyses of such data can only yield inferences about what was said, not *the intended meaning of* what was said.

Of course, utterances' intended meanings will only be of concern to researchers with a representational orientation. For example, Gottschalk's (1995) instrumental approach uses surface grammatical relations as symptoms of underlying psychological states, irrespective of how the source may have intended them. When from a representational orientation the researcher develops a semantic grammar to capture a specific type of linguistic expression (e.g., the grievance), text blocks must be scanned for instances of this type, which then, in turn, are encoded according to the grammar. In addition to such phenomenal semantic grammars, Roberts (1997b) has developed a generic semantic grammar for unambiguously encoding arbitrary clauses of natural language text. (For an application of the grammar, plus a symposium on the method see Roberts (1997c, 1997d) and Franzosi (1997b).)

Every network is a cluster of nodes and arcs. Consequently, network text analysts must also be careful to take into account the within-network clustering of the themes and theme-relations that respectively correspond to node- and node-arc-node-positions. Yet even more egregious errors may result if themes or theme-relations are themselves treated as units of analysis.

Note that one could conceivably generate a rather large data matrix from a single network if one were to create a separate row in the matrix for each of the network's themes or theme-relations, and to place in the cells of each row the values of variables measuring each of these themes' or theme-relations' various network-positions. From a statistical perspective, the consequences of such a strategy are extremely problematic. Not only are the new units of analysis (namely, the themes or theme-relations) not independent observations obtained via some random process, values on their associated variables are almost surely *dependent* observations with correlated errors. (For example, in a data matrix with one row per theme and with a column of conductivity-scores for each theme [viz., the number of linkages that the theme provides between other theme-pairs], dependence of scores across themes is ensured, because themes linked with highly conductive themes will, by virtue of this linkage, be more likely to be highly conductive themselves.) If these error correlations are not taken into account in one's analyses, the possibilities of severely biased estimates and of underestimated standard errors cannot be discounted.[7] On the other hand, there is a simple remedy: Restrict one's statistical analyses to data matrices with "the network" as unit of analysis. Errors will be uncorrelated if data on each row of one's data matrix is coded from a randomly sampled block of text.

Finally, in encoding text blocks as networks one must take into account that some types of links are transitive (e.g., if "A yields B" and "B yields C", then "A yields C") and others are not (e.g., if "A loves B" and "B loves C", then "A may not love C").[8] Claims that specific nodes are central, conductive, etc. will be vacuous unless these terms are understood as existing within a network having causal, equivalence, affective, or some other specific type of arc. When arc-types are ignored, one may only conclude that nodes are linked "in some unspecified way", leaving one with the same problems as are left to the classical text analyst who is unaware of the semantic relations between themes that co-occur.

6. Conclusion

Those who wish to draw statistical inferences about text-populations, will find themselves doing so based on a data matrix with text-related variables and, almost surely, contextual variables. Text-related variables in the matrix will measure occurrences of themes, theme-relations within a semantic grammar, and/or network-positions of themes and theme-relations. Possible contextual variables will indicate the source, message, channel, and/or audience uniquely associated with each text-block under analysis. Accordingly, quantitative text analysis can be used in answering questions about "what themes occur", "what semantic relations exist among the occurring themes", and "what network positions are occupied by such themes or theme relations" among texts with particular types of source, message, channel, or audience. Within these limits, the decision of which inferences to draw is, of course, where the researcher's own imagination must take hold.

Notes

[1] In contrast, qualitative text analysis methods are relatively more inductive, nonstatistical, and exploratory (cf., e.g., Berg, 1995: 2–4). Moreover, they can be combined with quantitative text analysis methods in the same study (Gray and Densten, 1998; Marshall and Rossman, 1995: 99–104).

[2] With this section I abandon the "Allerton House distinction" between representational and instrumental. The issue it raises, namely that speech acts may be devious or may otherwise not mean what they say (e.g., as in irony), is a problem requiring the coding of contextual variables

that describe what sorts of inauthenticities exist (Roberts, 1997b: 66–68). As I argue shortly, the distinction may certainly not be used in justifying applications of contingency analysis.

3. Note that classical text analysis is representational according to the Allerton House usage of the term, but instrumental according to Osgood's redefinition. Words and phrases are coded as "representative" instances of thematic categories (i.e., as having face validity according to common usage), yet the researcher uses data on these themes' occurrences to "instrumentally" diagnose their underlying meaning.
4. Although I have no tangible illustrations of instrumental network text analysis, one might imagine such an approach involving the "filling in" of networks according to the researcher's views on what network links (e.g., John reads pornography) are contingent on which (e.g., John hates women).
5. In this discussion I assume (without loss of generality) that the "sampling unit" and the "unit of analysis" are the same. With this assumption, one's units of analysis are identified as soon as one's text-population has been divided into distinct text-blocks. Yet more importantly, the assumption allows me to separate the problem at hand (i.e., the problem of identifying one's units of analysis), from problems related to the clustering of units of analysis within sampling units. I return to the clustering problem in the following section.
6. One PC-based statistical package that adjusts standard errors for such complicated sampling designs is PC CARP (The Survey Section, 219 Snedecor Hall, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.).
7. These problems have not gone unnoticed by researchers in social network analysis (Strauss and Ikeda, 1990, Walker et al., 1994; Wasserman and Faust, 1994). If network text analyses are of clusters of conditionally independent pairs of themes (possibly ones related in multiple ways), considerable statistical sophistication will be needed to ensure unbiased estimators with appropriately large standard errors.
8. Kleinnijenhuis et al. (1997) have done some innovative software development to ensure that inferences drawn from networks take into account the types of transitivity that may legitimately be applied across specific sequences of arcs.

References

Altheide, D. L. (1996). *Qualitative Media Analysis*. Thousand Oaks, CA: Sage.

Berg, B. L. (1995). *Qualitative Research Methods for the Social Sciences*. Boston: Bacon & Allyn.

Carley, K. M. (1986). An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* 12: 137–189.

Carley, K. M. (1988). Formalizing the social expert's knowledge. *Sociological Methods and Research* 17: 165–232.

Carley, K. M. (1997). Network text analysis: The network position of concepts. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 79–100.

Carley, K. M. & Palmquist, M. E. (1992). Extracting, representing, and analyzing mental models. *Social Forces* 70: 601–636.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cuilenburg, J. J. van, Kleinnijenhuis, J. & Ridder, J. A. de (1986). A theory of evaluative discourse: Towards a graph theory of journalistic texts. *European Journal of Communication* 1: 65–96.

Cuilenburg, J. J. van, Kleinnijenhuis, J. & Ridder, J. A. de (1988). Artificial intelligence and content analysis: Problems of and strategies for computer text analysis. *Quality & Quantity* 22: 65–97.

Denzin, N. K. & Lincoln, Y. S. (eds) (1994). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.

Eltinge, E. M. (1997). Assessing the portrayal of "science as a process of inquiry" in high school biology textbooks: An application of Linguistic Content Analysis. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 159–170.

Feldman, M. S. (1994). *Strategies for Interpreting Qualitative Data*. Thousand Oaks, CA: Sage.

Fielding, N. G. & Lee, R. M. (eds) (1991). *Using Computers in Qualitative Research*. London: Sage.

Franzosi, R. (1989). From words to numbers: A generalized and linguistics-based coding procedure for collecting event-data from newspapers. In: C. Clogg (ed.), *Sociological Methodology, 1989*. Oxford: Basil Blackwell, pp. 263–298.

Franzosi, R. (1990). Computer-assisted coding of textual data: An application to semantic grammars. *Sociological Methods and Research* 19: 225–257.

Franzosi, R. (1997a). Labor unrest in the Italian service sector: An application of semantic grammars. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 131–145.

Franzosi, R. (1997b). Comment: On ambiguity and rhetoric in (social) science. In: A. Raftery (ed.), *Sociological Methodology, 1997*. Oxford: Basil Blackwell, pp. 135–144.

George, A. (1959). Quantitative and qualitative approaches to content analysis. In: I. de S. Pool (ed.), *Trends in Content Analysis*. Urbana, IL: University of Illinois Press, pp. 7–32.

Gerner, D. J., Schrodt, P. A., Francisco, R. & Weddle, J. L. (1994). The analysis of political events using machine coded data. *International Studies Quarterly* 38: 91–119.

Goldhamer, D. H. (1969). Toward a more general Inquirer: Convergence of structure and context of meaning. In: G. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisley & P. J. Stone (eds), *The Analysis of Communication Content: Developments in Scientific Theories and Computer Techniques*. New York: Wiley, pp. 343–354.

Gottschalk, L. A. (1995). *Content Analysis of Verbal Behavior: New Findings and Computerized Clinical Applications*. Hillsdale, NJ: Lawrence Erlbaum.

Gottschalk, L. A. & Bechtel, R. (1989). Artificial intelligence and the computerization of the content analysis of natural language. *Artificial Intelligence in Medicine* 1: 131–137.

Gottschalk, L. A. & Kaplan, S. M. (1958). A quantitative method of estimating variations in intensity of a psychologic conflict or state. *Archives of Neurology and Psychiatry* 78: 656–664.

Gray, J. H. & Densten, I. L. (1998). Integrating quantitative and qualitative analysis using latent and manifest variables. *Quality & Quantity* 32: 419–431.

Griemas, A-J. (1984 [1966]). *Structural Semantics: An Attempt at a Method*. Lincoln, NE: University of Nebraska Press.

Grishman, R. (1986). *Computational Linguistics: An Introduction*. Cambridge, UK: Cambridge University Press.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*, 2nd edn. London: Arnold.

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

Honey, J. (1983). *The Language Trap: Race, Class and the 'Standard English' Issue in British Schools*. Harrow, UK: National Council for Academic Standards.

Kelle, U. (ed.) (1995). *Computer-Aided Qualitative Data Analysis: Theory, Methods, and Practice*. London: Sage.

Kleinnijenhuis, J., Ridder, J. A. de & Rietberg, E. M. (1997). Reasoning in economic discourse: An application of the network approach to the Dutch press. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 191–207.

Krueger, R. A. (1994). *Focus Groups: A Practical Guide for Applied Research*, 2nd edn. Thousand Oaks, CA: Sage.

Lasswell, H. D. (1948). The structure and function of communication in society. In: L. Bryson (ed.), *The Communication of Ideas*. New York: Harper & Row, pp. 37–51.

Lee, T.W. (1999). *Using Qualitative Methods in Organizational Research*. Thousand Oaks, CA: Sage.

Markoff, J. (1988). Allies and opponents: Nobility and the third estate in the spring of 1789. *American Sociological Review* 53: 477–496.

Markoff, J., Shapiro, G. & Weitman, S. (1974). Toward the integration of content analysis and general methodology. In: D. R. Heise (ed.), *Sociological Methodology, 1975*. San Francisco: Jossey-Bass, pp. 1–58.

Marshall, C. & Rossman, G. B. (1995). *Designing Qualitative Research*, 2nd edn. Thousand Oaks, CA: Sage.

McEnery, A. M. (1992). *Computational Linguistics: A Handbook and Toolbox for Natural Language Processing*. Wilmslow, UK: Sigma.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd edn. Thousand Oaks, CA: Sage.

Namenwirth, J. Z. & Weber, R. P. (1987). *Dynamics of Culture*. Winchester, MA: Allen & Unwin.

Osgood, C. E. (1959). The representational model and relevant research methods. In: I. de S. Pool (ed.), *Trends in Content Analysis*. Champaign, Ill: University of Illinois Press, pp. 33–88.

Pool, I. de S. (1955). *The Prestige Press: A Comparative Study of Political Symbols*. Cambridge, MA: MIT Press.

Pool, I. de S. (ed.) (1959). *Trends in Content Analysis*. Champaign, IL: University of Illinois Press.

Popping, R. (1997). Computer programs for the analysis of texts and transcripts. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 209–221.

Riessman, C. K. (1993). *Narrative Analysis*. Newbury Park, CA: Sage.

Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces* 68: 147–177.

Roberts, C. W. (1991). Linguistic content analysis. In: H. J. Helle (ed.), *Verstehen and Pragmatism: Essays on Interpretative Sociology*. Frankfurt: Peter Lang, pp. 283–309.

Roberts, C. W. (ed.) (1997a). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum.

Roberts, C. W. (1997b). Semantic text analysis: On the structure of linguistic ambiguity in ordinary discourse. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 55–77.

Roberts, C. W. (1997c). A generic semantic grammar for quantitative text analysis: Applications to East and West Berlin radio news content from 1979. In: A. Raftery (ed.), *Sociological Methodology, 1997*. Oxford: Basil Blackwell, pp. 89–129.

Roberts, C. W. (1997d). Reply: The curse of Chauvin. In: A. Raftery (ed.), *Sociological Methodology, 1997*. Oxford: Basil Blackwell, pp. 169–176.

Rosner, M. & Johnson, R. (eds.) (1992). *Computational Linguistics and Formal Semantics*. Cambridge, UK: Cambridge University Press.

Savaiano, S. & Schrodt, P. A. (1997). Environmental change and conflict: Analyzing the Ethiopian famine of 1984–85. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 147–158.

Schrodt, P. A. (1993). Machine coding of event data. In: R. L. Merritt, R. G. Muncaster & D. A. Zinnes (eds), *Theory and Management of International Event Data: DDIR Phase II*. Ann Arbor: University of Michigan Press, pp. 117–140.

Shapiro, G. (1997). The future of coders: Human judgments in a world of sophisticated software. In: C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 225–238.

Shapiro, G. & Markoff, J. (1998). *Revolutionary Demands: A Content Analysis of the Cahiers de Doléances of 1789*. Stanford: Stanford University Press.

Silverman, D. (1993). *Interpreting Qualitative Data: Methods for Analyzing Talk, Text, and Interaction*. London: Sage.

Smith, C. P. (ed.) (1992). *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge, UK: Cambridge University Press.

Stone, P. J., Dunphy, D. C., Smith, M. S. & Ogilvie, D. M. (eds) (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Strauss, D. & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association* 85: 204–212.

Walker, M. E., Wasserman, S. & Wellman, B. (1994). Statistical models for social support networks. In: S. Wasserman & J. Galaskiewicz (eds.), *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. Thousand Oaks, CA: Sage, pp. 53–78.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Weitzman, E. A. & Miles, M. B. (1995). *Computer Programs for Qualitative Data Analysis: A Software Sourcebook*. Thousand Oaks, CA: Sage.

Wolcott, H. F. (1994). *Transforming Qualitative Data: Description, Analysis, and Interpretation*, 2nd edn. Thousand Oaks, CA: Sage.