

# International Encyclopedia of the Social & Behavioral Sciences

*Editors-in-Chief*

Neil J. Smelser

*Center for Advanced Study in the Behavioral Sciences, Stanford, CA, USA*

Paul B. Baltes

*Max Planck Institute for Human Development, Berlin, Germany*

Volume 4



2001

ELSEVIER

AMSTERDAM—PARIS—NEW YORK—OXFORD—SHANNON—SINGAPORE—TOKYO

Elsevier Science Ltd., The Boulevard, Langford Lane, Kidlington, Oxford,  
OX5 1GB, UK

Copyright © 2001 Elsevier Science Ltd.

*All rights reserved. No part of this publication may be reproduced, stored in any retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic tape, mechanical, photocopying, recording or otherwise, without permission in writing from the publishers.*

First edition 2001

**Library of Congress Cataloging-in-Publication Data**

International encyclopedia of the social & behavioral sciences / editors in chief  
Neil J. Smelser, Paul B. Baltes. — 1st ed.

p. cm.

Includes bibliographical references.

ISBN 0-08-043076-7 (set : alk. paper)

I. Social sciences—Encyclopedias. I. Title: International encyclopedia of the  
social and behavioral sciences. II. Smelser, Neil J. III. Baltes, Paul B.

H41.158 2001

300'.3—dc21

2001044791

**British Library Cataloguing in Publication Data**

A catalogue record for this book is available from the British Library.

ISBN 0-08-043076-7 (set : alk. paper)

♻️ The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Typeset by Cambridge University Press, UK.

Printed and bound in Great Britain by Polestar Wheatons Ltd., Exeter, UK.

word. Whereas one might perform a content analysis of campaign slogans, one uses the data generated in such a content analysis to analyze (i.e., not 'content' analyze) the slogans' effects on public opinion. Texts and transcripts have been by far the most common objects of content analysis. For this reason, the focus of this article will be on the content analysis of texts (i.e., on *quantitative text analysis*). (For discussions of qualitative text analysis techniques, see *Linguistic Turn and Discourse Analysis in History; Event History Analysis: Applications; Ethnography; Qualitative Methods, History of*).

### 1. Classical Content Analysis

The first large-scale applications of content analysis came with the Second World War and US government commissioned propaganda analyses performed by Lasswell, Berelson, George, and other content analysis pioneers. During the first decade of the postwar era, a classical approach to content analysis emerged that involved the generation of data matrices of a particular type. These matrices typically exhibit a row for each sampled text block and a column for each theme (or meaning category) that may or may not have occurred in each block. Cells in the matrix are counts of the number of times the theme associated with its column occurred in the text block associated with its row. Thus, classical content analysis typically produces matrices of word- or phrase-counts, albeit sometimes ones embellished with value and salience weights (cf. Pool 1955, Stone et al. 1966, Holsti 1969, Smith 1992).

During a conference held in 1955 at the University of Illinois' Monticellos Allerton House many of these text-analysis pioneers gathered to develop solutions to the methodological problems of the day. Pool (1959) is the scholarly legacy of this conference. The most influential, if not the largest faction among the participants, was a group of Harvard researchers who made extensive use of what they called 'contingency analysis.' After generating a matrix of theme occurrences, these researchers would compute a matrix of associations between pairs of themes. Contingency analysis then proceeded as the researcher developed (usually *post hoc*) explanations of why some themes co-occurred (i.e., were positively associated) and why others were disassociated (i.e., were negatively associated).

In a casual but extremely insightful remark at the conference, Osgood (1959) noted, '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.'

## Content Analysis

Content analysis is a class of techniques for mapping symbolic data into a data matrix suitable for statistical analysis. When the term is used, one refers to the content analysis of cultural artifacts (e.g., books, architectural styles, discourse on prime-time television, etc.). That is, one refers to a mapping of non-numeric artifacts into a matrix of statistically manipulable symbols. Thus, content analysis involves measurement, not 'analysis' in the usual sense of the

In other words, classical content analysts treat words as symptomatic instruments from which the researcher can diagnose the source's possibly unconscious or unacknowledged characteristics.

## 2. Representational vs. Instrumental Interpretation

Shapiro (1997) has extended Osgood's meaning 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.' At issue in this distinction is 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 ... To 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). Accordingly, instrumental text analysis methods are used to identify individual and societal characteristics about which society's 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 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. In contrast, a researcher with an instrumental, Freudian orientation, would likely find the choice of matronly heroines to 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.

Methodological challenges related to representational text analysis are those associated with the researcher's ability to sympathetically understand the sources of his texts (see *Ethnography; Qualitative Methods, History of*). In both representational and instrumental text analyses the researcher is responsible

for developing an explicit method for systematically applying the sources' or researcher's perspectives in interpreting the texts under analysis.

## 3. Data Matrices Produced in Content Analyses of Texts

Independent of whether it is representational or instrumental, a content analysis always involves the production of a data matrix. That is, it requires that words be mapped into a two-dimensional matrix representation suitable for statistical analysis. This (i.e., content analysis's 'data matrix requirement') greatly simplifies the task of delineating the domain of possible questions that quantitative text analyses are able to address. In particular, it allows this 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. The balance of this article is devoted to further developing this observation, and consequently to helping the text analyst identify the most appropriate text analysis technique(s) for the research question at hand.

### 3.1 Variables

During the 1970s to the 1990s semantic and network text analysis methods have added to classical content analysis' word/phrase counts, other types of variables for the statistical analysis of linguistic data. Whereas in a thematic text analysis (of which classical content 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 inter-related. 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.

The type of data matrix generated in a thematic text analysis has already been described in the first paragraph of Sect. 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. 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 theme categories) can be constructed interactively (Popping 1997). 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 'in-

**Table 1**  
A data matrix from a semantic text analysis

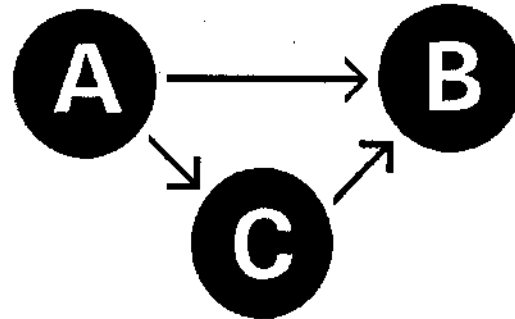
ID-number	Subject	Action	Object
1	10	34	53
2	8	21	78
3	13	34	61
4	10	35	61
5	16	29	44

strumental thematic text analyses' (also known as 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, e.g., 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 on semantic relations among themes such as 'political protest' and 'economic inflation' is not afforded by aggregated word-count data.

Relations among themes are encoded in a semantic text analysis, however. In a semantic text analysis the researcher begins by constructing a template (a semantic grammar). Themes from sampled texts are then mapped as syntactic components within this template. For example, Markoff et al. (1974) have developed a two-place semantic grammar for 'grievances' that contains 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 (1997) for more detailed descriptions of the application of semantic grammars in text analysis.

Table 1 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 1, 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 vs. 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 vs. whites refer to themselves as targets of aggression. More generally,



**Figure 1**  
A network of causal relations among themes.

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 to 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 (KEDS) is another special-purpose, parser-based program that can be used either instrumentally (in conjunction with the standard World Events Interaction Survey (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 (1997) research on labor disputes, Roberts's (1989) on ideology shifts, and Shapiro and Markoff's (1998) work on public opinion in eighteenth century France.

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.

For example, the simple three-theme network in Fig. 1 contains four sequences of causal linkages, namely  $A \rightarrow B$ ,  $A \rightarrow C$ ,  $C \rightarrow B$ , and  $A \rightarrow C \rightarrow B$ .

**Table 2**  
A data matrix from a network text analysis

ID-number	Causal salience measures					
	A on B	A on C	C on B	B on A	C on A	B on C
1	.50	.25	.25	.00	.00	.00
2	.25	.00	.50	.00	.25	.00
3	.00	.00	.25	.25	.50	.00
4	.00	.00	.00	.50	.25	.25
5	.00	.25	.00	.50	.00	.25
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Accordingly, the 'causal salience' of theme-A on theme-B equals 0.5, which is the proportion of the four sequences in which A is the cause and B is the effect. Note how calculation of this measure draws on more than isolated semantically-linked themes in a corresponding block of text. It incorporates information on all themes and links within the text-block's network representation.

Thus, a data matrix such as that in Table 2 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 pairs of other themes (Carley 1997). Another is of theme linkages that are logically implied, but not explicitly stated, in each block of text (Kleinnijhuis et al. 1997). In addition to pioneering work in this area by Osgood et al. (1956), nearly all quantitative network text analysis research has been conducted either by Carley or by a group of Dutch researchers in Amsterdam (Carley 1986, van Cuilenburg 1986, 1988, Carley and Palmquist 1992). All of these researchers do representational network text analysis. The dawn of instrumental network text analysis awaits a social scientist who both develops a network representation of language (analogous to a semantic grammar) and uses an interpretive perspective in developing rules for 'filling in' the network's arcs and nodes given various combinations of sentences in text blocks.

### 3.2 Units of Analysis

The discussion thus far has yielded a 2 x 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 content analysis of texts.

In the earliest stages of every quantitative text analysis, the researcher is confronted with a 'mountain of words' (a text-population) about which statistical inferences are to be drawn. On the one hand, this text population may be an initially undifferentiated mass (e.g., a sample of minutes of speech on US prime-time television could be drawn at random from the undifferentiated 'mountain of words' that were uttered during a year's 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.

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 divided initially.

Imagine a researcher, who wishes to analyze prime-time television data according to performers' mental models (or conceptual frameworks). In such a 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 short, 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 two types of units of analysis have been mentioned that might be yielded when a text population is divided; namely the conceptual framework of the text's source (e.g., the television performer's mental model), and the message that the text conveys (e.g., the program's story line). But of course, there are others.

Consider 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:

(a) characteristics of source (gender, affiliation, biases, etc.),

(b) characteristics of message (local vs. domestic news, descriptive vs. evaluative orientation, etc.),

(c) characteristics of channel (radio vs. TV news, public vs. commercial network, written vs. spoken medium, etc.), and

(d) characteristics of audience (sociocultural 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.

However, 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.

#### 4. What Questions Can Content Analyses Answer About Text Populations?

Content analyses of texts yield data matrices with text-related variables and, almost surely, contextual variables. Text-related variables in the matrix may measure occurrences of themes, theme-relations within a semantic grammar, and/or network-positions of themes and theme-relations. Contextual variables may indicate the source, message, channel, and/or audience uniquely associated with each text-block under analysis. Accordingly, content analyses of texts afford answers to 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 question to ask is, of course, where the researcher's own imagination takes hold.

*See also:* Interpretive Methods: Macromethods; Interpretive Methods: Micromethods

#### Bibliography

Carley K 1986 An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* 12: 137-89

Carley K M 1997 Network text analysis: The network position of concepts. In: Roberts C W (ed.) *Text Analysis for the Social*

*Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Lawrence Erlbaum, Mahwah, NJ, pp. 79-100

Carley K, Palmquist M 1992 Extracting, representing, and analyzing mental models. *Social Forces* 70: 601-36

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

van Cuilenburg J J, Kleinnijenhuis J, de Ridder J A 1988 Artificial intelligence and content analysis: problems of and strategies for computer text analysis. *Quality and Quantity* 22: 65-97

Franzosi R 1990 Computer-assisted coding of textual data: An application to semantic grammars. *Sociological Methods and Research* 19: 225-57

Franzosi R 1997 Labor unrest in the Italian service sector: An application of semantic grammars. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Lawrence Erlbaum, Mahwah, NJ, pp. 131-45

Gerner D J, Schrodt P A, Francisco R, Francisco R A, Weddle J L 1994 Machine coding of events using regional and international sources. *International Studies Quarterly* 38: 91-119

Gottschalk L A, Bechtel R 1989 Artificial intelligence and the computerization of the content analysis of natural language. *Artificial Intelligence in Medicine* 1: 131-37

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* 79: 688-96

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

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

Lasswell H D 1948 The structure and function of communication in society. In: Bryson L (ed.) *The Communication of Ideas*. Institute for Religious and Social Studies, New York, pp. 37-51

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

Namenwirth J Z, Weber R P 1987 *Dynamics of Culture*. Allen & Unwin, Boston

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

Osgood C E, Saporta S, Nunnally J C 1956 Evaluative assertion analysis. *Litera* 3: 47-102

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

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

Popping R 1997 Computer programs for the analysis of texts and transcripts. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Erlbaum, Mahwah, NJ, pp. 209-21

Roberts C W 1989 Other than counting words: A linguistic approach to content analysis. *Social Forces* 68: 147-77

## *Content Analysis*

---

- Roberts C W 1997 A generic semantic grammar for quantitative text analysis: Applications to East and West Berlin radio news content from 1979. *Sociological Methodology* 27: 89-129
- Shapiro G 1997 The future of coders: Human judgments in a world of sophisticated software. In: Roberts C W (ed.) *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Erlbaum, Mahwah, NJ, pp. 225-38
- Shapiro G, Markoff J 1998 *Revolutionary Demands: A Content Analysis of the Cahiers de Doléances of 1789*. Stanford University Press, Stanford, CA
- Smith C P (ed.) 1992 *Motivation and Personality: Handbook of Thematic Content Analysis*. Cambridge University Press, Cambridge, UK
- Stone P J, Dunphy D C, Smith M S, Ogilvie D M 1966 *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA

C. W. Roberts