

Third Edition

Content Analysis

An Introduction to Its Methodology

Klaus Krippendorff

The Annenberg School for Communication, University of Pennsylvania

 SAGE

Los Angeles • London • New Delhi • Singapore



Los Angeles | London | New Delhi
Singapore | Washington DC

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/11 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Acquisitions Editor: Matthew Byrne
Editorial Assistant: Stephanie Palermini
Production Editor: Astrid Virding
Copy Editor: Judy Selhorst
Typesetter: C&M Digitals (P) Ltd.
Proofreader: A. J. Sobczak
Indexer: Molly Hall
Cover Designer: Klaus Krippendorff
Marketing Manager: Liz Thornton
Permissions Editor: Karen Ehrmann

Copyright © 2013 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Krippendorff, Klaus.

Content analysis : an introduction to its methodology / Klaus Krippendorff.—3rd ed.

p. cm.

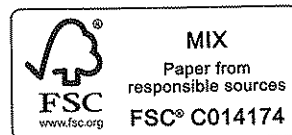
Includes bibliographical references and index.

ISBN 978-1-4129-8315-0 (pbk. : acid-free paper)
I. Content analysis (Communication) I. Title.

P93.K74 2013

401'.4--dc23 2011048278

This book is printed on acid-free paper.



12 13 14 15 16 10 9 8 7 6 5 4 3 2 1

Brief Contents

Preface to the Third Edition	xii
Introduction	1
PART I. CONCEPTUALIZING CONTENT ANALYSIS	
1. History	10
2. Conceptual Foundation	24
3. Uses and Inferences	49
PART II. COMPONENTS OF CONTENT ANALYSIS	
4. The Logic of Content Analysis Designs	82
5. Unitizing	98
6. Sampling	112
7. Recording/Coding	126
8. Data Languages	150
9. Analytical Constructs	170
PART III. ANALYTICAL PATHS AND EVALUATIVE TECHNIQUES	
10. Analytical/Representational Techniques	188
11. Computer Aids	208
12. Reliability	267
13. Validity	329
14. A Practical Guide	354
Glossary	380
References	391
Index	415
About the Author	441

Detailed Contents

Preface to the Third Edition	xii
Introduction	1
PART I. CONCEPTUALIZING CONTENT ANALYSIS	
1. History	10
1.1 Some Precursors	10
1.2 Quantitative Newspaper Analysis	11
1.3 Early Content Analysis	13
1.4 Propaganda Analysis	14
1.5 Content Analysis Generalized	17
1.6 Computer Text Analysis	19
1.7 Qualitative Approaches	22
2. Conceptual Foundation	24
2.1 Definition	24
2.2 Epistemological Elaborations	27
2.3 Examples	31
2.4 Framework	35
2.4.1 Texts	35
2.4.2 Research Questions	37
2.4.3 Context	38
2.4.4 Analytical Constructs	40
2.4.5 Inferences	41
2.4.6 Validating Evidence	44
2.5 Contrasts and Comparisons	45
3. Uses and Inferences	49
3.1 Traditional Overviews	49
3.2 Extrapolations	52
3.2.1 Trends	53
3.2.2 Patterns	54
3.2.3 Differences	55

3.3 Standards	58
3.3.1 Identifications	58
3.3.2 Evaluations	59
3.3.3 Judgments	61
3.4 Indices and Symptoms	61
3.5 Linguistic Re-Presentations	66
3.6 Conversations	70
3.7 Institutional Processes	71
3.8 Areas of Likely Success	77

PART II. COMPONENTS OF CONTENT ANALYSIS

4. The Logic of Content Analysis Designs	82
4.1 Content Analysis Designs	82
4.1.1 Components	84
4.1.2 Quantitative and Qualitative Content Analysis	88
4.2 Designs Preparatory to Content Analysis	90
4.2.1 Operationalizing Available Knowledge of the Context	90
4.2.2 Testing Analytical Constructs as Hypotheses	91
4.2.3 Developing a Discriminant Function	92
4.3 Designs Exceeding Content Analysis	94
4.3.1 Comparing Similar Phenomena Inferred From Different Bodies of Texts	94
4.3.2 Testing Relationships Among Phenomena Inferred From One Body of Texts	95
4.3.3 Testing Hypotheses Concerning How Content Analysis Results Relate to Other Variables	96
5. Unitizing	98
5.1 Units	98
5.2 Types of Units	99
5.2.1 Sampling Units	99
5.2.2 Recording/Coding Units	100
5.2.3 Context Units	101
5.3 Ways of Defining Units	104
5.3.1 Physical Distinctions	104
5.3.2 Syntactical Distinctions	105
5.3.3 Categorial Distinctions	106
5.3.4 Propositional Distinctions	106
5.3.5 Thematic Distinctions	108
5.4 Productivity, Efficiency, and Reliability	109
6. Sampling	112
6.1 Sampling in Theory	112
6.2 Sampling Techniques Applicable to Texts	114

6.2.1	Random Sampling	116
6.2.2	Systematic Sampling	116
6.2.3	Stratified Sampling	116
6.2.4	Varying Probability Sampling	117
6.2.5	Cluster Sampling	117
6.2.6	Snowball Sampling	118
6.2.7	Relevance Sampling	120
6.2.8	Census	121
6.2.9	Convenience Sampling	121
6.3	Sample Size	122
6.3.1	Statistical Sampling Theory	122
6.3.2	Sampling Experiments	123
6.3.3	The Split-Half Technique	125
7.	Recording/Coding	126
7.1	The Function of Recording and Coding	126
7.2	Coder Qualifications	128
7.2.1	Cognitive Abilities	128
7.2.2	Background	128
7.2.3	Frequency	129
7.3	Coder Training	129
7.4	Approaches to Defining the Semantics of Data	132
7.4.1	Verbal Designations	133
7.4.2	Extensional Lists	134
7.4.3	Decision Schemes	135
7.4.4	Magnitudes and Scales	136
7.4.5	Simulation of Hypothesis Testing	138
7.4.6	Simulation of Interviewing	139
7.4.7	Constructs for Closure	141
7.5	Records	142
7.5.1	Administrative Information	144
7.5.2	Information on the Organization of Records	144
7.5.3	Substantive Information About the Phenomena of Interest	146
8.	Data Languages	150
8.1	The Place of Data Languages in Analysis	150
8.2	Definitions	152
8.3	Variables	155
8.4	Nominal Variables	160
8.5	Orderings	161
8.5.1	Chains	161
8.5.2	Recursions	162
8.5.3	Cubes	163
8.5.4	Trees	163
8.6	Metrics	165

8.6.1	Ordinal Metrics	165
8.6.2	Interval Metrics	167
8.6.3	Ratio Metrics	168
8.7	Mathematical Operations	168
9.	Analytical Constructs	170
9.1	The Role of Analytical Constructs	170
9.2	Sources of Certainty	172
9.2.1	Previous Successes and Failures	172
9.2.2	Expert Knowledge and Experience	173
9.2.3	Established Theories	175
9.2.4	Embodied Practices	178
9.3	Types of Constructs	178
9.3.1	Extrapolations	178
9.3.2	Applications of Standards	179
9.3.3	Indices and Symptoms	179
9.3.4	Re-Presentations	181
9.3.5	Conversations/Interactions	181
9.3.6	Institutional Processes	182
9.4	Sources of Uncertainty	183
9.4.1	Variance of the Target	183
9.4.2	Confidence Levels	184
9.4.3	Appropriateness of the Construct	184
PART III. ANALYTICAL PATHS AND EVALUATIVE TECHNIQUES		
10.	Analytical/Representational Techniques	188
10.1	Counts	189
10.2	Cross-Tabulations, Associations, and Correlations	192
10.3	Multivariate Techniques	194
10.4	Factor Analysis and Multidimensional Scaling	198
10.5	Images, Portrayals, Semantic Nodes, and Profiles	199
10.6	Contingencies and Contingency Analysis	203
10.7	Clustering	205
11.	Computer Aids	208
11.1	What Computers Do	208
11.2	How Computers Can Aid Content Analyses	209
11.3	Text Analyses	214
11.3.1	Accounts of Character Strings	214
11.3.2	Text Mining	221
11.3.3	Mining Textual Evidence for Theories	229
11.3.4	Webgraph Analyses	232
11.4	Computational Content Analyses	235
11.4.1	Coding/Dictionary Approaches	239
11.4.2	Statistical Association Approaches	245

11.4.3	<i>Semantic Network Approaches</i>	248	14.	A Practical Guide	354
11.4.4	<i>Memetic Approaches</i>	254	14.1	Designing an Analysis	355
11.5	Qualitative Data Analysis Support	258	14.1.1	<i>Text-Driven Analyses</i>	355
11.6	Frontiers	262	14.1.2	<i>Problem-Driven Analyses</i>	357
11.6.1	<i>Intelligent Browsers</i>	263	14.1.3	<i>Method-Driven Analyses</i>	370
11.6.2	<i>Common Platforms</i>	263	14.2	Writing a Research Proposal	372
11.6.3	<i>Computational Theories of Meaning</i>	264	14.2.1	<i>Rhetorical Function</i>	372
11.6.4	<i>Utilization of Intertextualities</i>	264	14.2.2	<i>Contractual Function</i>	373
11.6.5	<i>Natural Interfaces</i>	265	14.2.3	<i>Outline for a Research Proposal</i>	374
12.	Reliability	267	14.3	Applying the Research Design	376
12.1	Why Reliability?	267	14.4	Narrating the Results	376
12.2	Reliability Designs	270	14.4.1	<i>Outline for a Research Report</i>	378
12.2.1	<i>Types of Reliability</i>	270			
12.2.2	<i>Conditions for Generating Reliability Data</i>	272	Glossary		380
12.2.3	<i>Reliability Data</i>	275	References		391
12.3	α -Agreement for Coding	277	Index		415
12.3.1	<i>The Direct Path to α</i>	280	About the Author		441
12.3.2	<i>The Indirect Path to α via Coincidence Matrices</i>	284			
12.3.3	<i>Difference Functions for Data With Various Metrics</i>	287			
12.3.4	<i>Some Typical Cases</i>	292			
12.3.5	<i>Some Contested Coefficients and Correspondences</i>	301			
12.4	α -Agreement for Unitizing	309			
12.4.1	<i>Reliability Data for Unitizing</i>	310			
12.4.2	<i>Difference Functions for Unitizing</i>	312			
12.4.3	<i>Computing α for Unitizing</i>	313			
12.4.4	<i>α-Agreement for Coding Identified Units of Unequal Size</i>	315			
12.5	Statistical Properties of α	319			
12.5.1	<i>Insufficient Variation</i>	319			
12.5.2	<i>Statistical Significance</i>	321			
12.5.3	<i>Sampling Considerations</i>	321			
12.5.4	<i>Standards for the Reliability of Data</i>	324			
13.	Validity	329			
13.1	Validity Defined	329			
13.2	A Typology for Validating Evidence	333			
13.2.1	<i>Sampling Validity</i>	336			
13.2.2	<i>Semantic Validity</i>	338			
13.2.3	<i>Structural Validity</i>	345			
13.2.4	<i>Functional Validity</i>	347			
13.2.5	<i>Correlative Validity</i>	348			
13.2.6	<i>Predictive Validity</i>	351			

Preface to the Third Edition

Content analysis is potentially one of the most important research techniques in the social sciences. The content analyst views data not as physical events but as texts, images, and expressions that are created to be seen, read, interpreted, and acted on for their meanings, and must therefore be analyzed with such uses in mind. Analyzing texts in the contexts of their uses distinguishes content analysis from other methods of inquiry.

Methods in the natural sciences are not concerned with meanings, contents, intentions, and references. Also, natural scientists hardly reflect on the textual sources of their own conceptions of nature, excluding them from their object of study, dismissing them as subjective in contrast to what can be determined through detached observation and objective measurement. Where social researchers adopt natural scientific methods of inquiry, the epistemology that is inscribed in such methods prevents them from addressing what matters most in everyday social life: human communication, how people coordinate their lives, the commitments they make to each other and to the conceptions of society they aspire to, what they know, and why they act. Certainly, content analysis is not the only research method that takes meanings seriously, but it is a method that has the additional qualities of being applicable to large numbers of data and being unobtrusive. It makes sense of what is mediated between people—textual matter, symbols, messages, information, mass-media content, and technology-supported social interactions—without perturbing or affecting those who handle that textual matter.

In the first edition of *Content Analysis*, published in 1980, I suggested that content analysis was at a crossroads. Content analysts at that time had a choice: They could continue their shallow counting game, motivated by a journalistic fascination with numbers and a narrow conception of science in which quantitative measurement provides the only evidence that counts (Lasswell, 1949/1965b), or they could redirect the attention of content analysts to social phenomena that are both generated by and constituted in texts and images and, hence, need to be understood through their written and pictorial constituents. I am pleased to say that the logic and methods that I presented in the first edition of *Content Analysis* have survived

the challenges provided by the radical transformation of the textual fabric of contemporary society, due in no small part to the ongoing information revolution. The increasingly widespread availability of electronic, and hence computer-readable, texts concerning virtually everything that matters to society and its members has moved content analysis, particularly computer-aided text analysis, into the center of how society examines itself.

In the 1980s, content analysis was a research method that had entered the psychological and social sciences but was used mainly by journalists and communication researchers. At that time, the amount of human effort required to collect, transcribe, and code textual data made content analysis a time-consuming and labor-intensive effort. Today, content analysis has become an efficient alternative to public opinion research, a method of tracking markets, political leanings, and emerging ideas; it is used as a way to settle legal disputes and as an approach to the exploration of individual human minds—not to dwell on the many improvements that content analysts have made in traditional content analytic inquiries of the mass media. Despite remarkable progress, content analysts can hardly claim to have met all challenges of this new era. Its potential is far from being exhausted by what can be done today, fueling the work of many developers of new analytic tools.

Although the outline of the second edition of *Content Analysis* (2004) remained essentially unchanged from that of the first, that edition clarified numerous methodological issues and responded to the technique's latest challenges. All chapters were substantially rewritten, addressing developments that had taken place since 1980, especially chapters that provided information on computer-aided text analysis and a practical guide that incorporated my own experiences in teaching and engaging in and consulting on academic and commercial research projects. Also, the earlier discussions of the epistemology, logic, and methods of content analysis were substantially revised. By comparison, this third edition introduces only minor updates, especially in the chapter on computer aids, and adds a glossary of terms.

I thank my students at the University of Pennsylvania's Annenberg School for Communication for their interest and for giving me useful feedback, and my colleagues for presenting me with the challenging methodological problems of their content analyses. I would also like to thank numerous readers of the earlier editions—both students and practicing content analysts—for sharing their comments and criticisms, and Sage Publications for its continuing support of content analysis literature, most recently by publishing *The Content Analysis Reader* (Krippendorff & Bock, 2009), which is meant to complement this edition with exemplary content analyses and accounts of the ways in which researchers have met various conceptual and methodological challenges.

The first edition of *Content Analysis* has been translated into Italian, Japanese, Spanish, and Hungarian, and during the 33 years since the publication of *Content Analysis* the book has reached an enormous audience. It has been widely adopted as a text in social science, humanities, and business curricula. It has served researchers as a guide to the design and execution of large and small content

analyses, and it has provided a standard for justifying as well as critically evaluating content analysis findings. When I travel to national and international conferences, I continue to be amazed and pleased to meet researchers from all over the world who tell me how studying this text has helped them in their current inquiries. In 2004, it received the International Communication Association (ICA) Fellows Book Award for its lasting contribution to communication research. This new edition is written for the same wide audience of practicing researchers, social scientists, and students.

—Klaus Krippendorff

*Gregory Bateson Professor Emeritus
for Cybernetics, Language, and Culture*

*The Annenberg School for Communication,
University of Pennsylvania*

Introduction

The term *content analysis* is about 60 years old. *Webster's Dictionary of the English Language* included the term in its 1961 edition, defining it as “analysis of the manifest and latent content of a body of communicated material (as a book or film) through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect.” The intellectual roots of content analysis, however, can be traced far back in human history, to the beginning of the conscious use of symbols and voice, especially writing. This conscious use, which replaced the magical use of language, has been shaped by the ancient disciplines of philosophy, rhetoric, and cryptography. It has also spawned religious inquisitions and political censorship on the part of ruling establishments. Today, symbolic phenomena are institutionalized in art, literature, education, and the mass media, including the internet. Theoretical and analytical concerns are found in such academic disciplines as anthropology, linguistics, social psychology, sociology of knowledge, and the comparatively younger field of communication studies. Many practical pursuits have grown from these fields: psychotherapy, advertising, politics, the arts, and so on. Virtually all disciplines within the whole spectrum of the humanities and the social sciences, including those that seek to improve the political and social conditions of life, are concerned with the functions and effects of symbols, meanings, and messages. In recent years, the emergence of the information society has moved the minutiae of communication—texts, contexts, images, interfaces, and, above all, information—into the very center of researchers’ attempts at self-understanding.

However ancient the roots of analyzing symbolic and textual matter might be, today’s content analysis is significantly different, in aim and in method, from that of the past. Contemporary content analysis has three distinguishing characteristics.

First, content analysis is an *empirically grounded method*, exploratory in process, and predictive or inferential in intent. Many of our current concepts relating to language are of Greek origin; for example, the words *sign*, *significance*, *symbol*, and *logic* all have Greek roots. However, the ancient Greeks’ interest in language was largely prescriptive and classificatory, not empirical. Aristotelian logic set the standards for clear expression, and much of rhetorical theory was directed toward a normative conception of persuasive argumentation. Science that explores rather than declares is a relatively recent accomplishment. Only a century ago, George

Boole and his contemporaries believed that the brain works according to (Boolean) logic and that human conduct is entirely rational. However, computers built on this logic turned out to be rather disappointing thinking machines. Empirical research in psychology is replacing Aristotelian categories in favor of a "psycho-logic." And we no longer measure human communication against the ideal of transmitting information. Instead, we inquire into what happens to the relationships between people who converse with one another.

With new conceptualizations and an empirical orientation, contemporary content analysts join other researchers in seeking valid knowledge or practical support for actions and critique. However, unlike researchers who employ other empirical techniques, content analysts examine data, printed matter, images, or sounds—texts—in order to understand what they mean to people, what they enable or prevent, and what the information conveyed by them does. These are questions for which natural scientists have no answers and for which their methods are generally insensitive.

Second, contemporary content analysis *transcends traditional notions of symbols, contents, and intents*. This may be seen in the evolution of the concept of communication, in how the development of media technologies has shaped our attention to communication, and in the role of culture in assigning significance to what is being analyzed. I would argue that in recent years our awareness of communication has undergone four conceptual revolutions, as described below, and probably is in the midst of a fifth:

- The idea of *messages*: the early awareness not only that verbal discourse is portable when written but also that writing has predictable effects. This awareness emerged in ancient Greece when messengers were used as the carriers of significance, history became documented, laws of the land were laid down in writing, and written instructions built organizational structures, directed events, and influenced (and possibly deceived) their receivers or the public. The concept of a message was a precursor of the rhetorical exploration of language. Tropes, syllogisms, and meanings came to be thought of as inherent qualities of speeches, letters, or documents. But a message is the metaphorical container of all these, a "container of content," a vehicle for shipping meanings from one place to another—for example, when we now leave a message for someone on voice mail or say that a message was meaningful (full of meanings) or meaningless (void of meanings).
- The idea of *channels*: the awareness of the constraints that every medium imposes on human communication. This awareness came with the increased reliance on different media of communication and served to explain their limitations: The alphabet limits what one can say in writing; the telephone confines communication to sound; and a television station can air no more than what is transmittable without interference from other stations, appealing to large audiences, and deemed profitable by its sponsors. The channel metaphor conjures images of canals and pipes with restricted capacities for shipping messages (with their contents) of certain forms and volumes.

- The idea of *communication*: the awareness of the relational space between senders and receivers, of the processes through which interpersonal relations are negotiated, social structures are constituted, and members of large populations come to know about each other. This awareness developed as an offshoot of the growth in mass media. By producing and disseminating identical messages—news and entertainment—to everyone, the mass media promised to be an agent of sharing, of building community relationships, of democratization, ideally, worldwide. Modeling themselves on the idea of mass production, the mass media also made us aware of where this one-way model failed: in interpersonal conversation, point-to-point telephone communication, public debate, and dialogue. In U.S. culture, mass-media technology has become synonymous with progress, and communication is understood as the cure for most social problems—for example, we often blame lack of communication or miscommunication when interpersonal as well as national conflicts arise.
- The idea of *systems*: the awareness of global, dynamic, and technologically supported interdependencies. This idea emerged with the growth of communication networks—telephone nets, wire services, mass-media systems, and most recently the internet—transforming commerce, politics, and interpersonal relationships, creating networks whose properties have so far defied attempts to theorize them adequately. Unlike the one-way mass media, systems are marked by the interactivity and simultaneity of parallel communication on a massive scale and with the potential of nearly universal participation.
- The idea of *computation*: the awareness of the algorithmic nature of certain routine cognitive and social processes and their increasingly sophisticated implementation in powerful computers. The processing of digital communications in place of print, visual media, and observed social practices, along with the ability to reproduce such data in visual and textual forms for reading, rearticulating, and disseminating by and to ideally everyone, is encouraging an entirely new literacy that undercuts traditional organizational structures, including national boundaries. The fluidity and enormous complexity that computation has introduced into almost all spheres of life amplify the possibilities for scientific exploration as well as present unprecedented challenges for collective understanding.

This rather sketchy history of communication suggests that researchers who are concerned with texts can no longer focus only on symbols or representations, nor can they limit themselves to questions about "who says what, through which channels, to whom, and with which effects" (Lasswell, 1960). The popular and simplistic notion of "content" has outlived its explanatory capabilities as well: *content*, the *what* of a communication, an entity that authors think they *enter* into messages and *ship* to remote receivers, who *remove* it for *what it is* and henceforth *share it* among others. This bizarre notion leads to authors as authorities of what they put into messages and to the conception of *content analysts* as experts who provide objective accounts of what messages were *intended to convey* or *actually contain*.

The virtuality of electronic media encourages short-lived access to messages that, without knowledge of their human authors, calls for a new technological basis for trust. It coordinates the lives of many people, overcoming old distinctions among channels of communication, obviating physical distances, and pushing capacities of the human participants to their limits. This erodes the validity of traditional communication theories, all the while enabling computer systems to thrive in this new environment. It is these computer systems that simulate and coordinate parts of the very social processes that researchers wish to understand. This is a radically changing world in which texts play distinctly new roles. Newspaper accounts, public opinion polls, corporate reports, files in government agencies, credit information, bank transactions, and, above all, huge textual data archives—all are now linked into networks that can be analyzed from numerous positions. In effect, the social systems that we conceived of as explaining society are now holographically retreating into our computers. This development calls for a redefinition of content analysis, one that aligns content—the target of the research—with how contemporary society operates and understands itself through its texts.

With the container metaphor rendered useless, perhaps the term *content analysis* no longer fits the realities of contemporary society. For better or for worse, I continue to use the term in this book, but I also plead with readers to oppose unflinchingly the naive and misleading entailments of the pervasive container metaphor.

Third, *contemporary content analysis has been forced to develop a methodology of its own*, one that enables researchers to plan, execute, communicate, reproduce, and critically evaluate their analyses whatever the particular results. Content analysts have had to develop such a methodology for three reasons:

- *Content analysts now face larger contexts.* The shift in interest from small collections of printed messages to systems and then to electronic texts and images circulating in the environment of content analysts is tied less to the nature of textual data than to the increasingly complex worlds that produce and are sustained by these data. This shift calls for theories and conceptions that earlier content analysts did not need. Although content analysts have frequently lamented the lack of general theories that could justify their work, progress in implementing more specific or micro-level theories is encouraging. This is especially true where content analysis has migrated through disciplines that were not previously concerned with textual data, such as the cognitive sciences and artificial intelligence.
- *Greater numbers of researchers need to collaborate in the pursuit of large-scale content analyses.* This observation is a correlate of the growing sample sizes of relevant texts, the analysis of which easily exceeds what individual analysts can handle. It implies that content analysts must work together, in parallel, and as research teams. Teamwork, however, needs to be organized reliably. Both the social problem of coordinating researchers and the methodological problem of assuring replicability tend to be solved through the adoption of a language whose vocabulary enables researchers to clarify the analytical

procedures they use, negotiate the individual responsibilities of the participants, assure agreement on the analytical categories, and evaluate the performance of team members.

- *The large volumes of electronically available data call for qualitatively different research techniques*, for computer aids. Such aids convert large bodies of electronic text into representations if not answers to research questions that content analysts need to understand. However, exactly what sophisticated text analysis software does—aside from promising to carry out the more labor-intensive clerical parts of processing textual data—is often difficult to retrace and inaccessible to the average content analyst. These computer aids participate in content analysis much as human analysts do. They become part of its methodology, with transparency being a major issue.

To be clear, *methodology* is not a value in itself. The purpose of methodology is to enable researchers to plan and examine critically the logic, composition, and protocols of research methods; to evaluate the performance of individual techniques; and to estimate the likelihood of particular research designs to contribute to knowledge. Every researcher must become proficient in defining the terms of an analysis and justifying the analytical steps taken to a skeptical friend or questioning colleague. Methodology provides a language for talking about the process of research, not about subject matter. In the history of scientific pursuits, the development of methodology has always been a major accomplishment. For example, for thousands of years humans preserved history by retelling or chanting stories, since the *Iliad* in writing, before the historian Leopold von Ranke, only a century ago, gave the “document” the methodological status it now has in the academic study of history. Similarly, scholars practiced “content analysis” well before Berelson and Lazarsfeld (1948) undertook the first codification of this method. Although many observers have argued that each content analysis is unique, possibly focusing largely on its subject matter, I would argue that all content analyses share a procedural logic and need to be justified through the use of socially acceptable criteria. These commonalities form the substance of this book.

I disagree with the frequent contention that content analysis is “nothing more than what everyone does when reading a newspaper, except on a larger scale.” Content analysis may have been that way, in its early, journalistic stage, and its methodology does not rule out such readings, but this narrow definition is no longer sufficient today. As newspaper readers, we are perfectly justified in applying our individual worldviews to texts and enacting our interest in what those texts mean to us; in fact, we cannot do otherwise. But as content analysis researchers, we must do our best to explicate what we are doing and describe how we derive our judgments, so that others—especially our critics—can replicate our results.

This book, then, introduces readers to ways of analyzing meaningful matter, texts, images, and voices—that is, data whose physical manifestations are secondary to what they mean to particular populations of people. The chapters are grouped into three main parts. Part I, “Conceptualizing Content Analysis,” begins with a brief chapter on the history of content analysis. In Chapter 2, I develop a definition

of content analysis that distinguishes this technique from other methods of inquiry, and in Chapter 3, I present a discussion of some of the ways in which content analysis has been applied. The chapters in Part II, "Components of Content Analysis," outline the procedures used in content analyses, beginning with their procedural logic and moving naturally from unitizing to sampling, recording/coding, data languages, and analytical constructs. The chapters in Part III, "Analytical Paths and Evaluative Techniques," trace several paths through content analysis protocols. In this part of the book, I discuss analytical constructs that enable researchers to draw inferences from data, the use of computers and computational techniques, and the two principal criteria used in evaluating content analyses: reliability and validity. In the final chapter, I provide a practical guide that summarizes the foregoing discussion from a practitioner's perspective.

Readers who have never done a content analysis may want to begin by reading Chapter 1, on the history of content analysis, and Chapter 3, on the uses of this technique, to get a sense for whether it suits their research interests. If it does, they should familiarize themselves with the conceptual foundations of content analysis by reading Chapter 2. Beginners in content analysis are advised to start with a small pilot project, to get a feel for what is involved in conducting a larger study. Methodology without some practice is empty. The guidelines in Chapter 14, although written as a summary, could also serve as a start. In this chapter, readers will find many helpful references to pertinent chapters in this volume, which may answer emerging questions and place these answers within the context of larger methodological issues. Beginning researchers will soon realize that analyzing text is not a mechanical task, and neither is designing a content analysis. Both undertakings require creativity and competence.

Readers who have had some experience with coding will acquire a larger perspective on what they had been doing. As the table of contents suggests, coding is only a small part of content analysis—despite popular misconceptions. In fact, only Chapter 7 is devoted to issues of coding or recording, something researchers need do only when their data or texts are unwieldy. By coding/recording textual matter, one learns to appreciate both the conceptual problems involved in imposing analytical categories on ordinary readings of text and the ways in which competent researchers have managed to solve such problems. Designing a content analysis is something different, however. I recommend that readers who have had experience with coding expand on that experience by examining the chapters offered here about all the other components of content analysis, adding these to their conceptual frameworks. Such readers might well look into Chapter 11, on computer aids, to gain an alternative perspective on manual unitizing and coding.

Readers who have already undertaken content analyses or similar text-based research will discover in this book alternative paths for such inquiries and a vocabulary that they can use in deliberating about what is involved in analyzing texts—not as observations of naturalistic phenomena but as data whose significance stems from the meanings that others bring to their readings. Those who think they know what content analysis is are advised to start with Chapter 2, on the conceptual foundations of content analysis. This chapter discusses the ways that researchers talk about content and exposes readers to the larger perspective they will need in order

to conceive a content analysis or critically evaluate the content analyses of others. As a condition for publication, scholarly journals increasingly demand some demonstration of why a content analysis should be taken seriously. In the past, content analysts relied heavily on conceptions of content as "contained" in messages, as discussed above, or "inherent" to texts. This settled the thorny issue of multiple text interpretations by fiat and consequently disabled explicitness about the researchers' procedures. Several research traditions—such as interpretive research, discourse analysis, literary scholarship, and rhetoric—tend to be plagued by similar conceptions. Researchers from these traditions would greatly benefit from explicating their approaches, checking their results against the work of others, and evaluating the social consequences of their findings outside their own schools of thought—as I am suggesting.

For experts in content analysis, this book raises several epistemological questions that practitioners rarely ask, transforms them into methodological ones, and provides new solutions to practical problems.

Readers who must make decisions concerning whether or not to trust the findings of content analyses and other text-based research—for instance, judges in courts of law, practitioners in the fields of public relations and advertising, and reviewers of research submitted for funding or publication in scientific journals—will find the vocabulary of this book useful as they need to weigh the quality of findings and make informed recommendations for improvements. Such readers will find the discussions in Chapters 2 (on the conceptual foundations), 12, and 13 (on reliability and validity) especially applicable to the necessary evaluative endeavors.

While this book may serve as a handbook for various practitioners, it grew out of my experiences in teaching courses and seminars in content analysis, and I conceive of it foremost as a textbook for advanced undergraduate and beginning graduate students. Teachers and their students may not want to work through all the chapters in their numerical order; for instance, those intending to use computers will find Chapter 11 more important than Chapter 7, on recording/coding, and may omit Chapter 12, on reliability, which is not a problem for software applications, but ought to consider the easily ignored validity of computer uses, discussed in Chapter 13. Students with specific projects in mind may pass over sections that may not be useful to their projects. However, readers should not rule out chapters as irrelevant before knowing the possibilities they offer.

Finally, for me, the book will have achieved its purpose if it helps to make the newly available wealth of electronic texts accessible to systematic analysis, if it improves the social significance of research in the humanities and the social sciences, and if it furthers the development of methods of inquiry into the realities we construct in processes of human communication.

PART I

Conceptualizing Content Analysis

CHAPTER 1

History

Empirical inquiries into the meanings of communications date back to theological studies in the late 1600s, when the Church found the printing of nonreligious materials to be a threat to its authority. Such inquiries have since mushroomed, moving into numerous areas and becoming the backbone of communication research. This chapter discusses several stages in the history of content analysis: quantitative studies of the press; propaganda analysis during World War II; social scientific uses of the technique in studies of political symbols, historical documents, anthropological data, and psychotherapeutic exchanges; computer text analysis and the new media; and qualitative challenges to content analysis.

1.1 Some Precursors

Content analysis entails a systematic reading of a body of texts, images, and symbolic matter, not necessary from an author's or user's perspective. Although the term *content analysis* did not appear in English until 1941 (Waples & Berelson, 1941, p. 2; cited in Berelson & Lazarsfeld, 1948), the systematic analysis of text can be traced back to inquisitorial pursuits by the Church in the 17th century. Religions have always been captivated by the written word, so it is not surprising that the first known dissertations about newspapers were defended in 1690, 1695, and 1699 by individuals pursuing academic degrees in theology. After the advent of the printing press, the Church became worried about the spread of printed matter of a nonreligious nature, and so it dealt with newspaper content in moralizing terms (Groth, 1948, p. 26). Surprisingly, in spite of the rhetorical tradition of ancient Greece, which was normative and oral in orientation, the 17th century contributed very little to the methodology of content analysis.

Probably the first well-documented quantitative analyses of printed matter occurred in 18th-century Sweden. According to Dovring's (1954–1955; see also

Krippendorff & Bock, 2009, Chapter 1.1) account, these analyses were undertaken as the result of the publication of the *Songs of Zion*, a collection of 90 hymns of unknown authorship. The collection had passed the Royal Swedish censor, but soon after its publication it was blamed for undermining the orthodox clergy of the Swedish state church. When the collection became popular, it was said to be "contagious" and was accused of aiding a dissenting group. Outstanding in this case is the fact that literary scholars of good reputation participated in the controversy, which crystallized around the question of whether the songs harbored dangerous ideas and, if so, how. Scholars on one side made a list of the religious symbols in the songs and became alarmed. Those on the other side, however, found the very same symbols in established songbooks and so discounted the claimed difference. Then some scholars noted that the symbols in the songs occurred in different contexts and had acquired meanings that were different from those taught in the official church. A debate arose about whether the meanings should be interpreted literally or metaphorically. The interpretations came to be compared with the results of a German study of the outlawed Moravian Brethren, a religious sect whose members later emigrated to the United States. This process—of revising a method in response to criticism—continued until it became clear to both sides in the debate how the symbols in the *Songs of Zion* differed from the symbols used in the official songbooks and how this (in the end political) phenomenon could be explained. The controversy generated many ideas that are now part of content analysis and stimulated debates about methodology that continue today.

In 1903, Eugen Löbl published in German an elaborate classification scheme for analyzing the "inner structure of content" according to the social functions that newspapers perform. His book, which became well-known in journalistic circles, contributed to the idea of *Publizistik*, or newspaper science, and foreshadowed functionalism, but it did not stimulate empirical investigations.

At the first meeting of the German Sociological Society in 1910, Max Weber (1911; see also Krippendorff & Bock, 2009, Chapter 1.2) proposed a large-scale content analysis of the press, but for a variety of reasons the research never got off the ground. During the same period, Andrei Markov (1913), who was working on a theory of chains of symbols, published a statistical analysis of a sample of Pushkin's novel in verse, *Eugene Onegin*. These inquiries were discovered only recently or influenced the content analysis literature only indirectly. For example, Weber is celebrated as one of the great sociologists, but his advocacy of the use of content analysis as a method for understanding the mass media is relatively unknown. And Markov's probability theories entered the content analysis literature only through Shannon's mathematical theory of communication (see Shannon & Weaver, 1949), which influenced Osgood's (1959) contingency analysis and cloze procedure.

Quantitative Newspaper Analysis 1.2

The beginning of the 20th century saw a visible increase in the mass production of newsprint. In the United States, the boom in newspapers created mass markets and interest in public opinion. Journalism schools emerged, leading to demands for

ethical standards and for empirical inquiries into the phenomenon of the newspaper. These demands, plus a somewhat simplistic notion of scientific objectivity, were met by what was then called *quantitative newspaper analysis*.

Probably the first quantitative newspaper analysis, published in 1893, asked the rhetorical question, “Do newspapers now give the news?” (Speed, 1893). Its author showed how, between 1881 and 1893, New York newspapers had dropped their coverage of religious, scientific, and literary matters in favor of gossip, sports, and scandals. In a similar but far more simplistic study published in 1910, Mathews attempted to reveal the overwhelming space that one New York daily newspaper devoted to “demoralizing,” “unwholesome,” and “trivial” matters as opposed to “worthwhile” news items. By simply measuring the column inches that newspapers devoted to particular subject matters, journalists in the early 20th century attempted to reveal “the truth about newspapers” (Street, 1909). Some believed that they had found a way of showing that the profit motive was the cause of “cheap yellow journalism” (Wilcox, 1900); others became convinced that they had established “the influence of newspaper presentations on the growth of crime and other antisocial activity” (Fenton, 1910). At least one concluded that a “quarter century survey of the press content shows demand for facts” (White, 1924).

Quantitative newspaper analysis seemingly provided the needed scientific ground for journalistic arguments. The respect for numbers has a long history, and facts that could be quantified were considered irrefutable. In a footnote, Berelson and Lazarsfeld (1948) quote from a source published more than 200 years ago:

Perhaps the spirit of the battle over ratification is best reflected in the creed ironically attributed to each of the contending parties by its opponents. The recipe for an Anti-Federalist essay which indicates in a very concise way the class-bias that actuated the opponents of the Constitution, ran in this manner: “wellborn, nine times—Aristocracy, eighteen times—Liberty of the Press, thirteen times repeated—Liberty of Conscience, once—Negro Slavery, once mentioned—Trial by Jury, seven times—Great men, six times repeated—Mr. Wilson, forty times . . .—put them together and dish them up at pleasure. (p. 9; quoted from *New Hampshire Spy*, November 30, 1787)

Quantitative newspaper analysis led to the development of many valuable ideas, however. In 1912, Tenney (see also Krippendorff & Bock, 2009, Chapter 1.4) made a far-reaching proposal for a large-scale and continuous survey of press content to establish a system of bookkeeping of the “social weather” “comparable in accuracy to the statistics of the U.S. Weather Bureau” (p. 896). He demonstrated what he had in mind with an analysis of a few New York newspapers for different ethnic groups, but his proposal exceeded the scope of what was then feasible. Quantitative newspaper analysis culminated in sociologist Malcolm M. Willey’s 1926 book *The Country Newspaper*. In this model study, Willey traced the emergence of Connecticut country weeklies, examining circulation figures, changes in subject matter, and the social role these papers acquired in competition with large city dailies.

When other mass media became prominent, researchers extended the approach first used in newspaper analysis—measuring volumes of coverage in various subject matter categories—initially to radio (Albig, 1938) and later to movies and television.

Content analysis in subject matter categories continues today and is applied to a wide variety of printed matter, such as textbooks, comic strips, speeches, and print advertising.

Early Content Analysis 1.3

The second phase in the intellectual growth of content analysis, which took place in the 1930s and 1940s, involved at least four factors:

- During the period following the 1929 economic crisis, numerous social and political problems emerged in the United States. Many Americans believed that the mass media were at least partially to blame for such problems as yellow journalism, rising crime rates, and the breakdown of cultural values.
- New and increasingly powerful electronic media of communication, first radio and later television, challenged the cultural hegemony of the newspapers. Researchers could not continue to treat these new media as extensions of newspapers, because they differed from the print media in important ways. For example, users of radio and television did not have to be able to read.
- Major political challenges to democracy were linked to the new mass media. For example, the rise of fascism was seen as nourished by the as-yet little-known properties of radio.
- Perhaps most important, this period saw the emergence of the behavioral and social sciences as well as increasing public acceptance of the theoretical propositions and empirical methods of inquiry associated with them.

In the 1930s, sociologists started to make extensive use of survey research and polling. The experience they gained in analyzing public opinion gave rise to the first serious consideration of methodological problems of content analysis, published by Woodward in a 1934 article titled “Quantitative Newspaper Analysis as a Technique of Opinion Research.” From writings about public opinion, interest in social stereotypes (Lippmann, 1922) entered the analysis of communications in various forms. Questions of representations were raised, with researchers examining topics such as how Negroes were presented in the Philadelphia press (Simpson, 1934); how U.S. textbooks described wars in which the United States had taken part, compared with textbooks published in countries that were former U.S. enemies (Walworth, 1938); and how nationalism was expressed in children’s books published in the United States, Great Britain, and other European countries (Martin, 1936).

One of the most important concepts that emerged in psychology during this time was the concept of “attitude.” It added evaluative dimensions to content analysis, such as “pro-con” or “favorable-unfavorable,” that had escaped the rough subject matter categories of quantitative newspaper analysis. Attitude measures redefined journalistic standards of fairness and balance and opened the door to the systematic assessment of bias. Among the explicit standards developed, Janis and Fadner’s (1943/1965) “coefficient of imbalance” deserves mention. Psychological experiments in rumor transmission led Allport and Faden to study newspaper content from an

entirely new perspective. In their 1940 article “The Psychology of Newspapers: Five Tentative Laws,” they attempted to account for the changes that information undergoes as it travels through an institution and finally appears on the printed page.

The interest in political symbols added another feature to the analysis of public messages. McDiarmid (1937), for example, examined 30 U.S. presidential inaugural addresses for symbols of national identity, of historical significance, of government, and of fact and expectations. Most important, Lasswell (1938), viewing public communications within his psychoanalytical theory of politics, classified symbols into such categories as “self” and “others” and forms of “indulgence” and “deprivation.” His symbol analysis led to his “World Attention Survey,” in which he compared trends in the frequencies with which prestige newspapers in several countries used national symbols (Lasswell, 1941; see also Krippendorff & Bock, 2009, Chapter 5.3).

Researchers in several disciplines examined the trends in scholarship, as reflected in the topics that representative journals published. Rainoff’s (1929) Russian study regarding physics was probably the first of this kind, but the most thorough analyses were conducted in the field of sociology (Becker, 1930, 1932; Shanas, 1945) and later in journalism (Tannenbaum & Greenberg, 1961).

Several factors influenced the transition from quantitative newspaper analysis, which was largely journalism driven, to content analysis:

- Eminent social scientists became involved in these debates and asked new kinds of questions.
- The concepts these social scientists developed were theoretically motivated, operationally defined, and fairly specific, and interest in stereotypes, styles, symbols, values, and propaganda devices began to replace interest in subject matter categories.
- Analysts began to employ new statistical tools borrowed from other disciplines, especially from survey research but also from experimental psychology.
- Content analysis data became part of larger research efforts (e.g., Lazarsfeld, Berelson, & Gaudet, 1948), and so content analysis no longer stood apart from other methods of inquiry.

The first concise presentation of these conceptual and methodological developments under the new umbrella term *content analysis* appeared in a 1948 mimeographed text titled *The Analysis of Communication Content*, authored by Berelson and Lazarsfeld, which was later published as Berelson’s *Content Analysis in Communications Research* (1952). This first systematic presentation codified the field for years to come.

1.4 Propaganda Analysis

Berelson described content analysis as the use of mass communications as data for testing scientific hypotheses and for evaluating journalistic practices. Yet the most important and large-scale challenge that content analysis faced came during World

War II, when it was employed in efforts to extract information from propaganda. Before the war, researchers analyzed texts in order to identify “propagandists,” to point fingers at individuals who were attempting to influence others through devious means. Fears concerning such influence had several origins. Propaganda was used extensively during World War I (Lasswell, 1927), and the years between the two world wars witnessed the effective use of propaganda by antidemocratic demagogues in Europe. In addition, Americans tend to have deep-seated negative attitudes toward religious fanatics, and the lack of knowledge concerning what the extensive use of the new mass media (radio, film, and television) could do to people raised concerns as well. According to the Institute for Propaganda Analysis (1937), propagandists reveal themselves through their use of tricks such as “name-calling,” employing “glittering generalities,” “plain folks” identifications, “card stacking,” “bandwagon” devices, and so on. Such devices could be identified easily in many religious and political speeches, even in academic lectures, and this approach to propaganda analysis led to a kind of witch-hunt for propagandists in the United States. Theories concerning subliminal messages, especially in advertising, raised widespread suspicion as well.

In the 1940s, as U.S. attention became increasingly devoted to the war effort, the identification of propagandists was no longer an issue. Nor were researchers particularly interested in revealing the power of the mass media of communication to mold public opinion; rather, military and political intelligence were needed. In this climate, two centers devoted to propaganda analysis emerged. Harold D. Lasswell and his associates, having written on political symbolism, worked with the Experimental Division for the Study of Wartime Communications at the U.S. Library of Congress, and Hans Speier, who had organized a research project on totalitarian communication at the New School for Social Research in New York, assembled a research team at the Foreign Broadcast Intelligence Service of the U.S. Federal Communications Commission (FCC). The Library of Congress group focused on analyzing newspapers and wire services from abroad and addressed basic issues of sampling, measurement problems, and the reliability and validity of content categories, continuing the tradition of early quantitative analysis of mass communications (Lasswell, Leites, & Associates, 1965).

The FCC group analyzed primarily domestic enemy broadcasts and surrounding conditions to understand and predict events within Nazi Germany and the other Axis countries, and to estimate the effects of Allied military actions on the war mood of enemy populations. The pressures of day-to-day reporting left the analysts little time to formalize their methods, and Berelson (1952) thus had little to say about the accomplishments of the FCC group. After the war, however, Alexander L. George worked through the volumes of reports that resulted from these wartime efforts to describe methods that had evolved in the process and to validate the inferences the researchers had made by comparing them with documentary evidence now available from Nazi archives. These efforts resulted in his book *Propaganda Analysis* (1959a; see also Krippendorff & Bock, 2009, Chapter 1.5), which made major contributions to the conceptualization of the aims and processes of content analysis.

The assumptions that propagandists are rational, in the sense that they follow their own propaganda theories in their choice of communications, and that the meanings of propagandists' communications may differ for different people reoriented the FCC analysts from a concept of "content as shared" (Berelson would later say "manifest") to conditions that could explain the motivations of particular communicators and the interests they might serve. The notion of "preparatory propaganda" became an especially useful key for the analysts in their effort to infer the intents of broadcasts with political content. In order to ensure popular support for planned military actions, the Axis leaders had to inform, emotionally arouse, and otherwise prepare their countrymen and -women to accept those actions; the FCC analysts discovered that they could learn a great deal about the enemy's intended actions by recognizing such preparatory efforts in the domestic press and broadcasts. They were able to predict several major military and political campaigns and to assess Nazi elites' perceptions of their situation, political changes within the Nazi governing group, and shifts in relations among Axis countries. Among the more outstanding predictions that British analysts were able to make was the date of deployment of German V weapons against Great Britain. The analysts monitored the speeches delivered by Nazi propagandist Joseph Goebbels and inferred from the content of those speeches what had interfered with the weapons' production and when. They then used this information to predict the launch date of the weapons, and their prediction was accurate within a few weeks.

Several lessons were learned from these applications of content analysis, including the following:

- Content is not inherent to communications. People typically differ in how they read texts. The intentions of the senders of broadcast messages may have little to do with how audience members hear those messages. Temporal orderings, individuals' needs and expectations, individuals' preferred discourses, and the social situations into which messages enter are all important in explaining what communications come to mean. Interpretations on which all communicators readily agree are rare, and such interpretations are usually relatively insignificant.
- Content analysts must predict or infer phenomena that they cannot observe at the time of their research. The inability to observe phenomena of interest tends to be the primary motivation for using content analysis. Whether the analyzed source has reasons to hide what the analyst desires to know (as in the case of an enemy during wartime or the case of someone needing to impress) or the phenomena of interest are inaccessible in principle (e.g., an individual's attitudes or state of mind, or historical events) or just plain difficult to assess otherwise (such as what certain mass-media audiences could learn from watching TV), the analyst seeks answers to questions that go outside a text. To be sure, the questions that a content analyst seeks to answer are the analyst's questions, and as such they are potentially at odds with whether others could answer them and how. Quantitative newspaper analysts made inferences without acknowledging their own conceptual contributions to

what they thought they found but actually inferred. Content is not the whole issue; rather, the issue is what can be legitimately inferred from available texts.

- In order to interpret given texts or make sense of the messages intercepted or gathered, content analysts need elaborate models of the systems in which those communications occur (or occurred). The propaganda analysts working during World War II constructed such models more or less explicitly. Whereas earlier content analysts had viewed mass-produced messages as inherently meaningful and analyzable unit by unit, the propaganda analysts succeeded only when they viewed the messages they analyzed in the context of the lives of the diverse people presumed to use those messages.
- For analysts seeking specific political information, quantitative indicators are extremely insensitive and shallow. Even where large amounts of quantitative data are available, as required for statistical analyses, these tend not to lead to the "most obvious" conclusions that political experts would draw from qualitative interpretations of textual data. Qualitative analyses can be systematic, reliable, and valid as well.

Convinced that content analysis does not need to be inferior to unsystematic explorations of communications, numerous writers in the postwar years, such as Kracauer (1947, 1952–1953) and George (1959a), challenged content analysts' simplistic reliance on counting qualitative data. Smythe (1954) called this reliance on counting an "immaturity of science" in which objectivity is confused with quantification. However, the proponents of the quantitative approach largely ignored the criticism. In his 1949 essay "Why Be Quantitative?" Lasswell (1949/1965b) continued to insist on the quantification of symbols as the sole basis of scientific insights. His approach to propaganda analysis produced several working papers but very few tangible results compared with the work of the FCC group of scholars. Today, quantification continues, although perhaps no longer exclusively.

Content Analysis Generalized 1.5

After World War II, and perhaps as the result of the first integrated picture of content analysis provided by Berelson (1952), the use of content analysis spread to numerous disciplines. This is not to say that content analysis emigrated from mass communication. In fact, the very "massiveness" of available communications continued to attract scholars who looked at the mass media from new perspectives. For example, Lasswell (1941) realized his earlier idea of a "world attention survey" in a large-scale study of political symbols in French, German, British, Russian, and U.S. elite press editorials and key policy speeches. He wanted to test the hypothesis that a "world revolution" had been in steady progress for some time (Lasswell, Lerner, & Pool, 1952). Gerbner and his colleagues pursued Gerbner's (1969) proposal to develop "cultural indicators" by analyzing, for almost two decades, one week of fictional television programming per year, mainly to establish "violence profiles" for different networks, to trace trends, and to see how various groups (such as

women, children, and the aged) were portrayed on U.S. television (see, e.g., Gerbner, Gross, Signorielli, Morgan, & Jackson-Beeck, 1979).

Psychologists began to use content analysis in four primary areas. The first was the inference of motivational, mental, or personality characteristics through the analysis of verbal records. This application started with Allport's (1942) treatise on the use of personal documents, Baldwin's (1942) application of "personal structure analysis" to cognitive structure, and White's (1947) value studies. These studies legitimated the use of written material, personal documents, and individual accounts of observed phenomena as an addition to the then-dominant experimental methods. A second application was the use of verbal data gathered in the form of answers to open-ended interview questions, focus group conversations, and verbal responses to various tests, including the construction of Thematic Apperception Test (TAT) stories. In the context of TAT stories, content analysis acquired the status of a supplementary technique. As such, it allowed researchers to utilize data that they could gather without imposing too much structure on subjects and to validate findings they had obtained through different techniques. Psychological researchers' third application of content analysis concerned processes of communication in which content is an integral part. For example, in his "interaction process analysis" of small group behavior, Bales (1950) used verbal exchanges as data through which to examine group processes. The fourth application took the form of the generalization of measures of meaning over a wide range of situations and cultures (which derived from individualist notions of meaning or content). Osgood (1974a, 1974b) and his students found numerous applications for Osgood, Suci, and Tannenbaum's (1957) semantic differential scales and conducted worldwide comparisons of cultural commonalities and differences.

Anthropologists, who started using content analysis techniques in their studies of myths, folktales, and riddles, have made many contributions to content analysis, including the componential analysis of kinship terminology (Goodenough, 1972). Ethnography emerged in anthropology, and although ethnographers often interact with their informants in ways that content analysts cannot interact with authors or readers, after ethnographers gather their field notes they start to rely heavily on methods that are similar to those that content analysts use.

Historians are naturally inclined to look for systematic ways to analyze historical documents, and they soon embraced content analysis as a suitable technique, especially where data are numerous and statistical accounts seem helpful. Social scientists also recognized the usefulness of educational materials, which had long been the focus of research. Such materials are a rich source of data on processes of reading (Flesch, 1948, 1951) as well as on a society's larger political, attitudinal, and value trends. In addition, literary scholars began to apply the newly available techniques of content analysis to the problem of identifying the authors of unsigned documents.

On one hand, this proliferation of the use of content analysis across disciplines resulted in a loss of focus: Everything seemed to be content analyzable; and every analysis of symbolic phenomena became a content analysis. On the other hand, this trend also broadened the scope of the technique to embrace what may well be the essence of human behavior: talk, conversation, and mediated communication.

In 1955, responding to increasing interest in the subject, the Social Science Research Council's Committee on Linguistics and Psychology sponsored a conference on content analysis. The participants came from such disciplines as psychology, political science, literature, history, anthropology, and linguistics. Their contributions to the conference were published in a volume titled *Trends in Content Analysis*, edited by Ithiel de Sola Pool (1959a). Despite obvious divergence among the contributors in their interests and approaches, Pool (1959a, p. 2) observed, there was considerable and often surprising convergence among them in two areas: They exhibited (a) a shift from analyzing the "content" of communications to drawing inferences about the antecedent conditions of communications and (b) an accompanying shift from measuring volumes of subject matter to counting simple frequencies of symbols, and then to relying on contingencies (co-occurrences).

Computer Text Analysis 1.6

The late 1950s witnessed considerable interest among researchers in mechanical translation, mechanical abstracting, and information retrieval systems. Computer languages suitable for literal data processing emerged, and scholarly journals started to devote attention to computer applications in psychology, the humanities, and the social sciences. The large volumes of written documents to be processed in content analysis and the repetitiveness of the coding involved made the computer a natural but also a difficult ally of the content analyst.

The development of software for literal (as opposed to numerical) data processing stimulated new areas of exploration, such as information retrieval, information systems, computational stylistics (Sedelow & Sedelow, 1966), computational linguistics, word processing technology, and computational content analysis. New software also revolutionized tedious literary work, such as indexing and the creation of concordances. Probably the first computer-aided content analysis was reported by Sebeok and Zeps (1958), who made use of simple information retrieval routines to analyze some 4,000 Cheremis folktales. In a Rand Corporation paper titled *Automatic Content Analysis*, Hays (1960) explored the possibility of designing a computer system for analyzing political documents. Unaware of both these developments, Stone and Bales, who were engaged in a study of themes in face-to-face interacting groups, designed and programmed the initial version of the General Inquirer system. This culminated in a groundbreaking book by Stone, Dunphy, Smith, and Ogilvie (1966) in which they presented an advanced version of this system and demonstrated its application in numerous areas, ranging from political science to advertising and from psychotherapy to literary analysis.

The use of computers in content analysis was also stimulated by developments in other fields. Scholars in psychology became interested in simulating human cognition (Abelson, 1963; Schank & Abelson, 1977). Newell and Simon (1963) developed a computer approach to (human) problem solving. Linguistics researchers developed numerous approaches to syntactic analysis and semantic interpretation of linguistic expressions. Researchers in the field of artificial intelligence focused on designing machines that could understand natural language (with very little success).

In 1967, the Annenberg School of Communications (which later became the Annenberg School for Communication) sponsored a major conference on content analysis. Discussions there focused on many areas—the difficulties of recording nonverbal (visual, vocal, and musical) communications, the need for standardized categories, the problems involved in drawing inferences, the roles of theories and analytical constructs, what developments content analysts could expect in the near future—but the subject of the use of computers in content analysis permeated much of the conference. Stone et al.'s (1966) book on the General Inquirer had just been published, and it had created considerable hope among content analysts. The contributions to the 1967 conference are summarized in a 1969 volume edited by Gerbner, Holsti, Krippendorff, Paisley, and Stone, the publication of which coincided with Holsti's (1969) survey of the field.

In 1974, participants in the Workshop on Content Analysis in the Social Sciences, held in Pisa, Italy, saw the development of suitable algorithms for computer content analysis as the only obstacle to better content analyses (Stone, 1975). Since that time, computational approaches have moved in numerous directions. One has been the development of customizable content analysis packages, of which the General Inquirer was the most important precursor. Attempts to apply the General Inquirer system to German texts revealed that software's English-language biases and led to more general versions of General Inquirers, such as TextPack. The basic ingredient of the General Inquirer and TextPack is a dictionary of relevant words. In the 1980s, Sedelow (1989) proposed the idea of using a thesaurus instead, as a thesaurus might be more accurate than a dictionary in reflecting "society's collective associative memory" (p. 4; see also Sedelow & Sedelow, 1986). In the 1990s, George Miller initiated a major research effort to chart the meanings of words using a computer-traceable network called WordNet (see Miller et al., 1993). In the 1980s, some authors observed that the enthusiasm associated with large systems that had appeared in the 1960s was fading (see Namenwirth & Weber, 1987), but today the development of text analysis software is proliferating, fueled largely by the historically unprecedented volumes of electronic and digital texts available for content analysis. More recently, Diefenbach (2001) reviewed the history of content analysis by focusing on four specific areas: mass communication research, political science, psychology, and literature.

Naturally, many researchers have compared computer-based content analyses with human-based content analyses. For example, Schnurr, Rosenberg, and Ozman (1992, 1993) compared the Thematic Apperception Test (Murray, 1943) with a computer content analysis of open-ended free speech and found the low agreement between the two to be discouraging. Zeldow and McAdams (1993) challenged Schnurr et al.'s conclusion, however. Nacos et al. (1991) compared humans' coding of political news coverage with data from Fan's (1988) computer-coded approach to the same coverage and found satisfactory correlations between the two. Nacos et al. came to the conclusion that content analysts can best use computers in their research by thinking of them as aids, not as replacements for the highly developed human capabilities of reading, transcribing, and translating written matter. As one might expect, today scholars hold many different opinions regarding the future of the use of computer-based content analysis.

Another development that has influenced how content analysts employ computers in their work is the increasingly common use of word processing software, which provides users with such features as spell-checkers, word- or phrase-finding and -replacing operations, and even readability indices. Although not intended for this purpose, ordinary word processing software makes it possible for a researcher to perform basic word counts and KWIC (keyword in context) analyses, albeit laboriously.

Word processing software is inherently interactive; it is driven by the user's reading of the textual material, not fixed. In the absence of computational theories of text interpretation, content analysts have found the symbiosis of the human ability to understand and interpret written documents and the computer's ability to scan large volumes of text systematically and reliably increasingly attractive. In such collaborations, human coders are no longer used as text-level content analysts; rather, they serve as translators of text or sections of text into categories that emerge during reading and then into a data language (that preserves relevant meanings), which enables various computational algorithms (that cannot respond to meanings) to do housekeeping and summarizing chores. This has given rise to a new class of software designed for computer-aided qualitative text analysis, of which NVivo and ATLAS.ti are two examples. Such interactive-hermeneutic text analysis software is becoming increasingly accessible, especially to students.

The most important stimulus in the development of computational content analysis, however, has been the growing availability of text in digital form. It is very costly to enter handwritten documents, such as transcripts of audio recordings of interviews, focus group protocols, minutes of business meetings, and political speeches, into a computer. Scanners have vastly improved in recent years, but they are still too unreliable to be used without additional manual editing. In the 1970s, data consortia emerged through which social scientists could share costly data, but the operations of these consortia were marred by a lack of standards and the usually highly specialized nature of the data. Then, in 1977, DeWeese proposed and took the remarkable step of bypassing the costly transcription process by feeding the typesetting tapes of a Detroit newspaper directly into a computer to conduct an analysis of the paper's content the day after it was published. Since that time, word processing software has come to be an integral part of the internal operations of virtually all social organizations; personnel create texts digitally before they appear on paper, use electronic mail systems, and surf the internet to download materials relevant to their work.

Today, a fantastic amount of raw textual data is being generated daily in digital form, representing almost every topic of interest to social scientists. Electronic full-text databases, to which all major U.S. newspapers, many social science and legal journals, and many corporations contribute all of the materials they publish, are growing exponentially and have become easily available and inexpensive to use online. Add to this the volume of electronic publications, the research potential of the internet, data available from online multiuser discussions (MUDs) and newsgroups, and online survey systems, which may well replace focus groups and interviews in certain empirical domains, and it is clear that the landscape of how society presents itself has been altered drastically. With more and more people interested in

this wealth of digital data, there is a corresponding demand for increasingly powerful search engines, suitable computational tools, text base managing software, encryption systems, devices for monitoring electronic data flows, and translation software, all of which will eventually benefit the development of computer-aided content analysis. The current culture of computation is moving content analysis into a promising future.

1.7 Qualitative Approaches

Perhaps in response to the now dated “quantitative newspaper analysis” of more than a century ago or as a form of compensation for the sometimes shallow results reported by the content analysts of 60 years ago, a variety of research approaches have begun to emerge that call themselves *qualitative*. I question the validity and usefulness of the distinction between quantitative and qualitative content analyses. Ultimately, all reading of texts is qualitative, even when certain characteristics of a text are later converted into numbers. The fact that computers process great volumes of text in a very short time and represent these volumes in ways someone can understand does not remove the qualitative nature of the texts being analyzed and the algorithms used to process them: On the most basic level, computers recognize zeros and ones and change them as instructed, proceeding one step at a time. Nevertheless, proponents of qualitative approaches to content analysis offer alternative protocols for exploring texts systematically.

Discourse analysis is one such approach. Generally, *discourse* is defined as text above the level of sentences. Discourse analysts tend to focus on how particular phenomena are represented. For example, Van Dijk (1991) studied manifestations of racism in the press: how minorities appear, how ethnic conflicts are described, and how stereotypes permeate given accounts, for example, in advertisements during sports events (Wonsek, 1992). Other discourse analysts have examined how television news programs and other TV shows in the United States manifest a particular ideological vision of the U.S. economy (Jensen, 2006), the components of “age markers” in the humorous context of the TV series *The Golden Girls* (Harwood & Giles, 1992), and the portrayal of the peace movement in news editorials during the Gulf War (Hackett & Zhao, 1994).

Researchers who conduct *social constructivist analyses* focus on discourse as well, but less to criticize (mis)representations than to understand how reality comes to be constituted in human interactions and in language, including written text (Gergen, 1985). Such analysts may address how emotions are conceptualized (Averill, 1985) or how facts are constructed (Fleck, 1935/1979; Latour & Woolgar, 1986), or they may explore changing notions of self (Gergen, 1991) or of sexuality (Katz, 1995).

Rhetorical analysis, in contrast, focuses on how messages are delivered, and with what (intended or actual) effects. Researchers who take this approach rely on the identification of structural elements, tropes, styles of argumentation, speech acts, and the like; Kathleen Hall Jamieson’s book *Packaging the Presidency* (1984) is an example of such an analysis. Efforts to study negotiations (Harris, 1996), what works and what doesn’t, might be described as rhetorical analyses as well.

Ethnographic content analysis, an approach advocated by Altheide (1987), does not avoid quantification but encourages content analysis accounts to emerge from readings of texts. This approach works with categories as well as with narrative descriptions but focuses on situations, settings, styles, images, meanings, and nuances presumed to be recognizable by the human actors/speakers involved.

Conversation analysis is another approach that is considered to be qualitative. The researcher performing such an analysis tends to start with the recording of verbal interactions in natural settings and aims at analyzing the transcripts as records of conversational moves toward a collaborative construction of conversations. This tradition is indebted to the work of Harvey Sacks, who studied numerous interactive phenomena, including the collaboration among communicators in the telling of jokes (Sacks, 1974). Goodwin (1977, 1981) extended conversation analysis by incorporating video data in his groundbreaking study of turn taking.

Qualitative approaches to content analysis have their roots in literary theory, the social sciences (symbolic interactionism, ethnomethodology), and critical scholarship (Marxist approaches, British cultural studies, feminist theory). Sometimes they are given the label *interpretive*. They share the following characteristics:

- They require a close reading of relatively small amounts of textual matter.
- They involve the rearticulation (interpretation) of given texts into new (analytical, deconstructive, emancipatory, or critical) narratives accepted within particular scholarly communities that are sometimes opposed to positivist traditions of inquiry.
- The analysts acknowledge working within hermeneutic circles in which their own socially or culturally conditioned understandings constitutively participate. (For this reason, I refer to these approaches as *interactive-hermeneutic*, a description that speaks to the process of engaging in systematic interpretations of text.)

One could summarize and say that content analysis has evolved into a repertoire of methods of research that promise to yield inferences from all kinds of verbal, pictorial, symbolic, and communication data. Beyond the technique’s initially journalistic roots, the past century has witnessed the migration of content analysis into various fields and the clarification of many methodological issues. After a short period of stagnation in the 1970s, content analysis is today growing exponentially, largely due to the widespread use of computers for all kinds of text processing. As of February 2011, an internet search for “*content analysis*” using the Google search engine found 1,650,000 documents. By comparison, “*survey research*” turned up 275,000 hits and “*psychological testing*,” 894,000. Since the first casual mention of “content analysis” in 1941—that is, seventy years ago and with a frequency of one—the public interest in the body of content analysis research has clearly grown to an astonishing extent.