Extending the Utility of Content Analysis

via the Scientific Method

Kimberly A. Neuendorf, Ph.D.

and

Paul D. Skalski, Ph.D.

School of Communication

Cleveland State University

Cleveland, OH 44115

216-687-3994

k.neuendorf@comcast.net

Manuscript in support of presentation to the Social Science and Social Computing Workshop, University of Hawaii, Honolulu, HI, May 23, 2010

Extending the Utility of Content Analysis via the Scientific Method

Introduction

Although quantitative content analysis has a history nearly as long as institutionalized survey research (Rogers, 1994), rigorous methodological standards have not always been enforced. Content analysis has been the "poor sister" of quantitative methods, notably with regard to the standards of validity and reliability (Lombard et al., 2002; 2004; Neuendorf, 2009; Pasadeos, Huhman, Standley, & Wilson, 1995). Even contemporary reviews of content analyses find such salient standards as reliability assessment to be lacking in a majority of published studies of human coding. For example, a recent systematic analysis of 133 health media content analyses (Neuendorf, 2009) found not a single instance of full reliability assessment and reportage, with 38% including no reliability assessment whatsoever (comparable to the 31% figure found by Lombard et al., 2002, in their review of content analysis in the field of communication).

Thus, this paper will first lay the groundwork for standards that must be achieved if content analysis is to be deemed a worthy set of procedures for inquiry and application to analysis of social networks and other messages, using characteristics of the scientific method as a guide. Throughout, a practical approach is taken.

Content Analysis Defined

First, a definition of content analysis is in order, to establish a common understanding of methodological assumptions. A range of definitions of content analysis have been articulated, from Babbie's broad net (”the study of recorded human communications," 2010, p. G2) to definitions such as Weber's, that assume the ability to make facile inferences ("a research method that uses a set of procedures to make valid inferences from text," 1990, p. 9). A more restrictive definition of content analysis is adopted here: Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method, including attention to objectivity/intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing. It is not limited as to the type of messages that may be analyzed, nor as to the types of constructs that might be measured (Neuendorf, 2002, p. 10).

Two main content analysis methodological choices exist—Human coding, and computer coding (i.e., computer-aided text analysis, CATA; for fuller coverage of CATA, see Gottschalk & Bechtel, 2008; Neuendorf, 2002; Roberts, 1997; West, 2001)[i]. While the main interest of scholars reading this paper might immediately fall in the realm of CATA, it is important to understand the vital contribution of human coding techniques for (a) the origination of content analytic schemes that eventually become CATA algorithms, (b) the measurement of constructs that are highly latent[ii] (and, correspondingly, for which a line of research has not yet devised adequate CATA indicators), and (c) the ongoing validation of CATA measures, as language standards and practices (idioms, etc.) evolve over time.

As noted, within the realm of quantitative content analysis, rigorous standards have not always been met. While the norms vary by journal (Neuendorf, 2011), and, fortunately, the quality bar seems to be rising over time, a gap still exists between the rigor required for an acceptable (i.e., publishable) content analysis and that required for other quantitative methods such as survey or experimental techniques. A number of scholars with wide experience in content analysis and other quantitative research approaches have attempted to close this gap (Krippendorff, 2004; Lombard et al., 2002;

Neuendorf, 2002; Riffe, Lacy, & Fico, 2005). A core recommendation, then, is that all content analyses should be guided by accepted methodological sources that are informed by an array of methodological and applied research experiences, and optimally not limited to a single discipline.

That said, the above definition assumes a quantitative approach and subscription to the tenets of the scientific method, with its goal of generalizable knowledge and its concomitant functions of description, prediction, explanation, and control (Hanna, 1969; Kaplan, 1964). While scholarship focused on message analysis needs to be committed to the use of a variety of methodologies, this paper will detail the particular needs of the quantitative content analysis researcher whose aim it is to describe, predict, explain, and control (ultimately to reproduce) human communication phenomena.

Focus on CATA (Computer-Aided Content Analysis)

CATA began with the General Inquirer, a mainframe computer application introduced by Philip Stone of Harvard in 1965 (Stone et al., 1966). The purpose of the General Inquirer was to automate the analysis of textual information, searching for text that delineates such features as valence, Osgood’s three semantic dimensions, language reflecting particular institutions, emotion-laden words, cognitive orientation, and more. All CATA programs have as their basis the analysis of text via the application of some algorithms of word or word sequence searching and counting. Most often, the analysis involves one or more dictionaries, i.e., lists of search terms intended to measure constructs on the text. Tables 1 and 2 provide summary information about 10 commonly-used CATA programs that range from those that provide only word-count output for researcher-created dictionaries, to those that include multiple pre-set dictionaries (i.e., dictionaries that are part of the CATA package), to those that allow no dictionaries but instead focus on patterns of co-occurrences of words.

Two prominent CATA programs that include well-documented pre-set dictionaries are LIWC and Diction 5.0. In LIWC (Linguistic Inquiry and Word Count; Pennebaker, Booth, & Francis, 2007) there are 84 dictionaries that tap such linguistic and semantic concepts as use of first-person pronouns, anger, optimism, reference to home, and reference to motion. The program Diction 5.0 (Hart, 2000), designed to analyze political speech, has 31 pre-set dictionaries, including those intended to measure tenacity, aggression, praise, satisfaction, and complexity. The 31 dictionaries are also combined to form “master variable” scales: Activity, optimism, certainty, realism, and commonality. The alternative to using pre-set dictionaries is to create one’s own custom dictionaries, and most CATA programs allow for this. However, the development of original dictionaries is quite demanding and ought to include a validation process that links measured dictionaries with additional indicators of the construct under investigation.

Content Analysis and Science

In order for the scientific basis of content analysis to be fully realized, there needs to be substantially enhanced corroboration and consultation among scholars and practitioners across interdisciplinary boundaries. For example, instances of computer text analysis of online content are sometimes presented from an engineering point of view, which is not well understood by behavioral scientists. Other significant advancements in the study of online communication are being made by psychologists, communication scientists, and linguists. These literatures stand almost completely separated from one another, a state of affairs that impedes mutual progress. Importantly, the sheer terminology may require cross-disciplinary translation (e.g., Abbasi, Chen, Thoms, Fu, 2008). For example, in an explicit attempt to bridge human coding and computer coding in a study of discussion threads in a secondary school online forum, Lin, Hsieh, and Chuang (2009) refer to the validation of their automated genre classification system (GCS) against human coding as "coherence," making it difficult to spot as a counterpart to traditional scientific notions of "measurement validity." And, what is termed "sentiment analysis" in key contemporary research literatures (e.g., Liu, 2010) has its clear conceptual and methodological roots in the branch of psychometrics that uses text analysis to infer psychological states and motivations from naturally-occurring speech (e.g., Lieberman, 2008; Lieberman & Goldstein, 2006; Smith, 1992).

At the root of certain key cross-disciplinary differences is a varying emphasis on the mutually-informative processes of induction and deduction. As a scientific enterprise, content analysis is held to the standard of deductive testing of hypotheses. However, much critical information is available from more inductively based processes such as text mining and the similar word-pattern discovery of such applications as CATPAC. Optimally, text structures derived from CATPAC analyses may help build theoretic models that then may be tested via deductive processes.

The goals of deductive hypothesis testing and (more broadly) scientific theory include description, prediction, explanation, and control, each of which will now be discussed in more detail as they relate to content analysis.

Description. The most fundamental scientific function, that of description, is readily achieved in a variety of fashions through human coding or CATA. In particular, the measurement of text via CATA techniques can generate such descriptives as:

(a) Occurrence (frequency)—of words, average word length, average sentence length (factored into readability);

(b) Context of occurrence—concordances (e.g., Bird et al., 2009), or keyword-in-context (KWIC);

(c) Location of occurrence—e.g., dispersion plots in Python (a natural language processing program), indicating visually where specific language appears within a text;

(d) Co-occurrence—type/token ratios, collocations (Bird et al., 2009, p. 20), dimensional analyses (including cognitive maps) that may reveal multiple domains of discourse (e.g., using such softwares as CATPAC); this last technique has been used in realms as diverse as tourist destination-image research and Czech activists' online communication (Stepchenkova, Kirilenko, & Morrison, 2009 and Hajek & Kabele, 2010, respectively);

(e) Networks—a type of co-occurrence summarization, network analysis examines the interaction patterns among communicative nodes (Mei, 2008). A very specific, applied example is the PostHistory and Social Network Fragments visual interfaces (Viegas et al., 2004), developed as means of displaying an emailer's network of contacts either (a) over time in a vertical timeline, or (b) in a map that indicates topical proximities (via physical space) and strengths of ties (via font size).

Prediction. A more complex and challenging function of science is that of prediction. For many writers in the philosophy of science literature, it is the veritable Holy Grail, reified in contemporary times by authors such as Imre Lakatos. According to

Lakatos (1978), the hallmark of science is its ability to make novel predictions. Predictions separate progressive from degenerative programs of research. In progressive research programs, predictions can successfully be made, such as Halley's prediction that a comet would return based on Newtonian theory. Prediction therefore serves as a demarcation criterion distinguishing science from pseudoscience, in addition to having tremendous practical value.

While prediction is certainly a more powerful status than description, prediction without explanation may be a hollow victory. Without explanatory ability, the relationships and predictions remain atheoretic. For example, Naccarato and Neuendorf (1998) achieved significant statistical prediction of print ad recall, with 59% of the variance "explained" by a set of content analysis variables (human coded) selected via stepwise inclusion. However, the set of significant predictors was not isomorphic with what prevailing theoretic literature would support. Thus, the researchers were left with the ability to predict readership within a narrowly proscribed set of conditions (e.g., business-to-business print ads within the electric power industry) but were not able to give a full explanation as to why the prediction worked, and why other, equally reasonable variables failed to enter into the prediction.

In another example, Bird, Klein, and Loper (2009) have noted that effective machine translation may be achieved via "collecting massive quantities of parallel texts from news and government websites that publish documents in two or more languages" (p. 30) and then establishing correspondences that may be applied to novel examples of text in one of those languages. Thus, a valid translation "prediction" may be achieved without any real knowledge of the substance of the text example.

Kucukyilmaz et al. (2008) contend that features of online messages may be used to predict both user- and message-specific attributes, including the biological, social, and psychological attributes of the source (p. 1453). In their chat mining analysis (2008), they included predictive variables from 10 stylistic feature categories, including word length, types of punctuation used, usage of stopwords and smileys, and vocabulary richness. Their statistical prediction was able to distinguish authorship, source gender, internet domain origin (.edu vs. .com vs. .net), and even day vs. night message origin (although which specific variables were included in the various statistical models is unclear).

Content analyses—both human-coded and CATA—have advanced measures for psychometrics, both in general measurement contexts (termed “thematic content analysis”) and clinical applications. For all psychometrics, the goal is to infer source characteristics from message attributes, including both substance and form.

Computer-driven linguistic analyses of texts have been used to create profiles of deceptive language (Newman, Pennebaker, Berry, & Richards, 2003; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004). Such studies have produced four general classifications of cues that are associated with deception: number of words, (b) types of pronouns used, (c) use of emotion words, and (d) indicators of cognitive complexity (Hancock, Curry, Goorha, & Woodworth, 2008, p. 3). Experimental work by Zhou, Burgoon, Nunamaker, and Twitchell (2004) found significant differences between subject-generated truthful and deceptive CMC analyzed via an automated Linguistic Based Cues system, providing a set of factors that may be useful as predictors of naturally-occurring online deception. Predictors of deception included a greater number of words, greater informality, lower lexical diversity, and use of more modifiers.

In these studies and others, the goal has been to "predict" from messages to source characteristics. This type of inference was first promised by Berelson (1952) and reinforced by Krippendorff (1980). However, the ready inference to source from message content has been contested, and scholars have recognized the value of integrating studies of message content, sources, receivers, and message effects (Shoemaker & Reese, 1996). Neuendorf (2002) proposes an "integrative" approach to content analysis, whereby source, contextual, and/or receiver data are collated with message data, providing solid relationship paths that will provide an empirical basis for future inferences.

If the paths are well-worn (well established via replication), we may be able to infer back to source characteristics from language attributes. This is what Chung and Pennebaker (2007; 2008) and others have worked to establish over the past decade. The promise of Berelson (1952) is bearing fruit through these studies—we are more able, reliably and validly, to infer characteristics of sources from types and patterns of language use by those sources.

Explanation. Perhaps the highest level of scientific achievement is explanation—i.e., a thorough understanding of the underlying mechanisms driving relationships among variables (Salmon, 1999). This understanding is integral to the development and testing of theory and an important distinction from prediction alone. Some scholars have honed in on the separability of prediction and explanation (e.g., Kaplan, 1964). At a certain point in the history of science, prediction and explanation were viewed as symmetric—

two sides of the same conceptual process, as it were (Hanson, 1959; Hempel, 1965). This symmetry thesis is composed of two parts—every successful explanation is a potential prediction, and every successful prediction is a potential explanation (Ruben, 1990, p. 146), although the thesis, particularly the latter part, has been hotly debated (Angel, 1967; Koertge, 1991; Rescher, 1997; Salmon, 1999). Fitting the former part of the thesis, sociologist Kurt Lewin is famously quoted as saying, “There is nothing so practical as a good theory.”

There has been an unfortunate divorce of explanation and prediction in the philosophy of science literature in recent decades; as Douglas (2009) notes, each can profit from a consideration of the other: “[A]ccounts of explanation have been impoverished by the neglect of prediction. . . explanation should be understood as a cognitive tool that assists us in generating new predictions. This view of explanation and prediction clarifies what makes an explanation scientific and why inference to the best explanation makes sense in science” (2009, abstract).

Establishing explanation is a challenge, requiring the isolation of many factors, often over a long period of multiple research tests. The explanatory mechanisms underlying important predictions are not always easily discernible. For example, McClelland, Koestner, and Weinberger (1992) found text analysis measures of achievement motive more useful (predictive) indicators for assessing learning than self-report measures, a counterintuitive and problematic finding for the attempt to establish a full explanatory model.

Control. Control may be effected whether or not explanation has been achieved. Prediction may be enough. For example, if there are paths of adequate strength established between source characteristics and message attributes, then artificial language production that evinces desired source characteristics (e.g., happiness, anger, humor) is feasible. Oshita (2010) has applied information about relationships between semantic text content and appropriate physical motion to generate appropriate computer animation from natural language texts, with no information about the intermediate processes needed to produce desired outcomes.

Standards

Successful performance of these four functions of science is reliant on achieving the standards of scientific inquiry. As with any systematic empirical investigation, a content analysis should proceed only after adequate planning and preparation. As outlined elsewhere (Neuendorf, 2011), six points summarize the major decision elements faced by the content analyst: (1) Theoretic and conceptual backing--each content analysis must be guided by a theoretic framework, either as a direct test of theory or utilizing theory primarily as an underlying rationale for the study of messages; (2) A plan for the scope of the investigation--a determination of whether the particular study will include only analyses of message content/form, or integrate these data with source and/or receiver data (Neuendorf, 2002; Shoemaker & Reese, 1996)[iii]; (3) A review of past research in anticipation of the development of content analysis measures—i.e., a human coding scheme or CATA dictionary set; (4) Defining the population of messages--the set of units (in content analysis, messages or message components) to which the researcher wishes to generalize; (5) Immersion in the message pool (immersion may disclose key variables that might otherwise go unrecognized); (6) Deciding whether to use human coding and/or computer coding (CATA)--a wide variety of computer programs now provide pre-set dictionaries intended to measure such constructs as optimism, aggression, and emotional tone (Neuendorf & Skalski, 2009), and most allow the development of custom dictionaries by the researcher (although the latter is a conceptual and logistic challenge).

Reliability and Validity

With human coding, a critical concern is that of reliability, the stability of a measure across human raters, time, and other conditions. Much concern has been invested in the pursuit of raising the bar with regard to this standard (Neuendorf, 2009). Rigorous coder training, and intercoder and intracoder reliability checks with appropriate statistical assessment, must be conducted and fully reported.

With computer coding (i.e., CATA), the concern shifts to measurement validity and its assessment. Clearly, validity is also important for human-coded schemes, but unfortunately the quest for reliability typically seems to overshadow it. With the automaticity of CATA, reliability is a given, while the process raises questions of validity.

When using CATA, decisions must obviously be made regarding how to establish dictionaries and other search and measurement criteria. More than a dozen quantitative CATA programs are available, and most include some pre-set dictionaries. Even though the validity of CATA programs ought to be scrutinized, few CATA procedures have been subjected to validation processes of any type. In fact, the same type of construct validity assessment that is traditionally demanded of survey and experimental measures is rarely employed in content analysis, human or computer (McAdams & Zeldow, 1993; Short et al., 2010). Those that have include LIWC, as well as Gottschalk and Bechtel’s PCAD

2000 (Psychiatric Content Analysis and Diagnosis) (2002; 2008), a computer application that analyzes naturally occurring speech in order to provide psychiatric diagnostic guidelines for a number of clinical classifications derived from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (e.g., anxiety disorders, schizophrenia, depression, and cognitive impairment stemming from dementia or substance abuse). Following a long history of validation of an earlier human coding scheme (Gottschalk & Gleser, 1969), Gottschalk and Bechtel's CATA program has been validated repeatedly against expert clinical diagnoses.

In the best case, validation may be part of the very development protocol for the CATA system. Lin, Hsieh, and Chuang (2009) developed an automatic text coding procedure for the identification of the "genre" of an online discussion thread; the computer coding compared favorably with expert judges' assessments.

Oddly, one type of validation possible for thematic content analysis has rarely appeared in the literature—a check of affect or attitude derived from text against self-report measures of affect. This seemingly obvious type of criterion validation is routinely ignored in favor of expert judge assessments. However, McClelland, Koestner, and Weinberger's (1992) efforts to correlate self-assessments of the achievement motive with those derived from coded stories and pictures found only moderate relationships, with coded findings more predictive of learning outcomes than self-report. This points to the possibility of CATA measures providing a window to a communicator's psychological orientations that is not available via self-report, due either to the individual's unwillingness to provide a valid report or the individual's lack of awareness of their status.

Recently, Short, Broberg, Cogliser, and Brigham (2010) have provided a comprehensive illustration of a model for validity assessment of CATA. They discuss how CATA as a content analysis technique may be improved through a rigorous consideration of construct validation procedures, including tests for content validity, external validity, dimensionality, and predictive validity (plus reliability). They identify specific steps that can be taken by the content analyst in each of these areas. For example, content validity, or the extent to which a measure samples the entire content domain of a construct, may be established through a combination of deductive and inductive methods. Short et al. recommend constructing CATA dictionaries by first using theory or a conceptual definition to generate a list of terms representing a construct, in deductive fashion. They then suggest taking the inductive step of generating a word frequency list from preliminary content and having multiple coders independently choose other words representing the construct of interest, with their judgments subjected to an intercoder reliability test. The inductive words can be then be added to the deductive terms to provide a thorough representation of the content area for actual coding.

Short et al. make other valuable linkages between scientific validation procedures and CATA considerations as well. To establish external validity or generalizability of content, they suggest comparing results across multiple sampling frames. To assess dimensionality, they recommend keeping word lists representing a multi-dimensional construct separate and then conducting tests on results to see if subdimensions should remain separate or be collapsed. And to establish predictive validity, the authors suggest relating results to dependent variables of interest not captured by content analysis.

To illustrate these steps, the article culminates with an application of the validation procedures to the concept of “entrepreneurial orientation.” At the content validity stage, the authors first generated six dictionaries (autonomy, innovativeness, proactiveness, competitive aggressiveness, risk taking, and an additional inductive dictionary) to represent the construct. They then sampled from both S & P 500 firms and Russell 2000 stock index firms to provide a comparison coding group for external validity assessment. For dimensionality, the authors compared results across the six dictionaries and two sample frames in a correlation matrix. Finally, they related these variables to measures of firm performance to establish predictive validity, using multiple regression analyses.

In the discussion, Short et al. examine other potential validation considerations (such as a consideration of discriminant validity) and the utility of supplementing machine coding with human coding as a way to further validate constructs. Overall, this article does a noteworthy job of linking CATA/content analysis with aspects of the scientific method. It focuses strictly on the use of CATA/content analysis in research on organizational communication but the recommendations can easily be extended to other areas of inquiry.

Generalizability

An overarching goal of science is to produce generalizable knowledge. At its essence, content analysis seeks nomothetic, not idiographic, knowledge--It is not aimed at an understanding of a single communicative message, but rather at the study of multiple messages so as to draw conclusions about messages, or possibly about the sources or cultural contexts of those messages.[iv]

At the core of the notion of generalizability is the application of measures to a set of units (messages, in the case of content analysis) that is representative of a population of interest. When the researcher needs to select a subset of units from the population, probability (random) sampling is essential if generalization to the defined, larger population of messages is desired. However, a valid sampling frame that enumerates the entire population is not always available, and the use of such nonrandom techniques as convenience sampling, purposive sampling, or quota sampling might be necessary. The size of the sample should be established with accepted statistical practices (Riffe, Lacy, & Fico, 2005).

The particular medium in which the messages are carried will clearly affect the sampling process (e.g., availability of sampling frame, units of sampling) as it does the population definition. For example, for content analyses of web sites, it is typical that a “snapshot” approach is used for collecting the sample (Norris, 2003). For example, Curtin and Gaither (2003) downloaded entire web sites, collecting their content twice, one month apart, in order to capture the “dynamic nature of the web” (p. 12). This freezing of the content is essential to reliability.

Application: Social Networks and other Online Communication

Some scholars contend that the online environment provides a particularly valid and neutral locus for unfettered, natural communication. And, it is the locus of a huge volume of digitized message content, generally highly accessible. These factors make online communication a fruitful context for content analyses that may contribute to the construction of predictive and explanatory scientific models.

Following early fears of a limited “cues filtered out” model of communication (Sproull & Kiesler, 1986; Walther & Parks, 2002), scholars began to recognize the ways in which the “so-called impediments of communication technology are overcome by its users” (Walther, 2004, p. 386). The “hyperpersonal” model stands in contrast to the limited model, suggesting that computer-mediated communication (CMC) may actually facilitate social interaction because communicators may take more time and greater care in creating messages than they would face-to-face (FtF) (Duthler, 2006; Walther, 2007).

Newton, Kramer, and McIntosh’s (2009) CATA (i.e., LIWC) study of blogging by individuals with autism spectrum disorders (ASP) found similarities in language choice between ASP and neurotypical bloggers that suggest the communication deficits exhibited by ASP communicators in face-to-face contexts are due to social contextual cues. The asynchronous nature of CMC may offer the user a “place of refuge” for communicative expression, and, as noted by Kim et al. (2007), online communication may facilitate communication for individuals who are shy or might otherwise be marginalized in FtF interactions.

As noted earlier, new modes of communication may engender new linguistic and communication norms. Research on computer-mediated communication (CMC) reveals striking differences in how online text differs from non-CMC naturally occurring writing, as summarized by Abbasi and Chen (2008): CMC is richer in interaction (both synchronous and asynchronous forms), CMC is less topical, and CMC technologies allow the emergence of novel language varieties (p. 813). Recognizing the evolving nature of online language, Neviarouskaya, Prendinger, and Ishizuka (2007; 2009) have supplemented natural language processing techniques with the inclusion of web-centric "symbolic cues" such as emoticons and abbreviations in their analyses of source affect/emotion/sentiment as expressed through online writings.

Content analyses of CMC at first focused on web site content, blogs, and other postings intended for a generalized online audience.[v] CATA analyses have focused on online discussion threads in an educational context (Lin, Hsieh, & Chuang, 2009; Pena-Shaff & Nicholls, 2004) and academic articles from online databases (Hu et al., 2009). An important set of studies of online messages has attempted to infer affect (or emotion or sentiment) from the text of web forums and blogs (Abbasi, Chen, Thoms, & Fu, 2008), therapeutic support bulletin boards (Lieberman & Goldstein, 2006), online news (Tian & Stewart, 2005), and personal blogs (Neviarouskaya, Prendinger, & Ishizuka, 2009).

Other research has approached online content with a stylometric focus, identifying authorship (Abbasi, Chen, & Nunamaker, 2008; Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2008) in an automated fashion, extending the long tradition of authorship attribution work by trained or expert coders. Similarly, Opoku, Pitt, and Abratt (2007) applied Aaker's brand personality theory to bestselling authors' official web sites, finding that five personality constructs measured by newly-created dictionaries (sincerity, excitement, competence, sophistication, and ruggedness) resulted in correspondence analysis mapping that clearly differentiated among the authors.

Paralleling their dramatic rise in use in the past five years, online social networks have been the subject of an increasing number of content analyses investigating a variety of topics. These include several studies focusing specifically on MySpace, including content analyses of personal information disclosure and communication on MySpace (Jones, Millermaier, Goya-Martinez, & Schuler, 2008), self-presentation on MySpace

(Kane, 2008), cultural differences in posting of MySpace comments (Lunk, 2008), predictors of user suicide on MySpace profiles (Kobayashi, Spitzberg, & Anderson, 2008), relational motivation and age effects on MySpace self disclosure (Jinsuk, Klautke, & Serota, 2009), and an analysis of adolescent use of MySpace over time (Patchin & Hinduja, 2010). This last study documented the abandonment of MySpace by many users over the past few years in favor of Facebook and other social networking sites, which have been studied (along with MySpace) in more recent content analyses. These include investigations of the use of social networking sites by candidates in the 2008 Presidential election (Compton, 2008), speech acts in social networking site users' status messages (Carr, Shrock, & Dauterman, 2009), ethnic and racial identity displays in Facebook profiles (Grasmuck, Martin, & Shanyang, 2009), and the use of Facebook by non-profit organizations (Waters, Burnett, Lamb, & Lucas, 2009). The wide variety of studies conducted on social networking sites in a relatively short period of time illustrates the rich possibilities these sites present for empirical inquiry.

One troubling feature these studies all share, however, is a reliance on human coding rather than CATA (or a combination of the methods) to get at features of interest (Kane, 2008; Lunk, 2008). CATA has not been attempted much (if at all) in studies of online social networks, and a perusal of popular sites suggests possible reasons why. First, online social networking sites are an amalgamation of static text *plus* chat, pictures, audio, video, and other types of content not suited to traditional CATA techniques. To further complicate matters, these features are now presented in a highly complex, idiosyncratic fashion. For example, Facebook now hides long discussions, personal information, and other content of interest behind tabs and buttons on a member's profile.

To view them, users have to click on what they want to see. Researchers using traditional CATA programs may have a difficult time accessing this information, unless researchers physically extract the content and put it into a format compatible with a CATA program.

Despite these and other challenges, there are some promising future possibilities for successfully using CATA to study social networking and other more complex online content. Web mining, web crawlers, Internet bots, and related technologies have demonstrated the feasibility of having automated computers analyze and even interact with web content of all types. The increasing capability of computers to "surf" the web themselves are the reason why users sometimes have to enter difficult-to-read text from images before registering for sites or making online purchases, for example. Computers have limitations but have shown a dramatic increase in what they can do over time. New CATA programs may have to be written (or existing ones updated) to extract certain content on popular online social networks, but this capability seems to exist already.

In a study illustrating the potential for extracting and analyzing online content, Kucukyilmaz, Cambazoglu, Aykanat, and Can (2008) used CATA to predict the identities of 100 online authors from their chat logs, with 99.7% accuracy. They concluded that creators of online content have distinct communication styles and word selection habits that can be detected. If the full potential of this type of automated content analysis realized, there are obvious benefits in a number of application areas (e.g., e-commerce, identifying security threats), but limitations still exist, as shown in the study by Kucukyilmaz et al. First, the external validity of this work is questionable, since it only looked at chat logs in Turkish language. It remains to be seen if these findings would be replicated with other forms of online communication, and with other languages.

Second, the chat data were extracted from a website set up for the study, which recorded the chat messages. How can chat messages be sampled from popular chat applications (e.g., IM and Facebook) that are not set up to archive conversations? Third, Kucukyilmaz et al. had to use a complex combination of CATA plus some human intervention to obtain prediction at a high level. It remains to be seen if this can be done through CATA alone. Nevertheless, the Kucukyilmaz et al. research supports the established finding that communicator characteristics such as word selection and linguistic style can be used for predictive purposes. It also calls attention to this potential with novel forms of communication (such as chat) in an online environment.

Conclusion

The challenges to producing useful applications of content analysis that employ the standards of scientific inquiry, particularly within the context of automated analyses, are substantial. However, the benefits of a scientific approach are even more substantial, including greater confidence in knowledge and the ability to predict future outcomes. It may be tempting to just let a computer program generate output, but the knowledge created through such a purely automated approach will not be as useful or make as much sense unless scientific standards are adhered to. As Carl Sagan famously stated, science is a “candle in the dark” (Sagan, 1997). Science sheds light on empirical phenomena and makes visible findings even clearer. We argue that use of CATA and human coding for research purposes should always be guided by scientific principles.

References

Abbasi, A., & Chen, H. (2007). Categorization and analysis of text in computer mediated communication archives using visualization. *Proceedings of the 7th ACM/IEE Joint Conference on Digital Libraries—Building & Sustaining the Digital Environment*, 11-18.

Abbasi, A., & Chen, H. (2008). CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, *32*, 811-837.

Abbasi, A., Chen, H., & Nunamaker, J. F. Jr. (2008). Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, *25*(1), 49-78.

Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *20*, 1168-1180.

Angel, R. B. (1967). Explanation and prediction: A plea for reason. *Philosophy of Science*, *34*(3), 276-282.

Babbie, E. R. (2010). *The practice of social research* (12th ed.). Belmont, CA: Wadsworth Cengage.

Berelson, B. (1952). *Content analysis in communication research*. New York: Hafner.

Binsted, K., Bergen, B., Coulsen, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, D., & O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, *21*(2), 59-69.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media, Inc.

Carr, C., Schrock, D., & Dauterman, P. (2009). Speech Act Analysis Within Social Network Sites' Status Messages. *Conference Papers -- International Communication Association*, 1-38. Retrieved from Communication & Mass Media Complete database.

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, *42*, 96-132.

Chung, C. K., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social Communication* (pp. 343-359). New York: Psychology Press.

Compton, J. (2008). Mixing Friends with Politics: A Functional Analysis of '08 Presidential Candidates Social Networking Profiles. *Conference Papers -- National Communication Association*, 1.

Curtin, P., & Gaither, K. (2003). *Public relations and propaganda in cyberspace: A quantitative content analysis of Middle Eastern government websites*. Paper presented at the annual meeting of the International Communication Association, San Diego, CA.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*, 444-463.

Duthler, K. (2006). The politeness of requests made via email and voicemail: Support for the Hyperpersonal Model. *Journal of Computer-Mediated Communication, 11*, 500-521.

Franiuk, R., Seefelt, J. L., & Vandello, J. A. (2008). Prevalence of rape myths in headlines and their effects on attitudes toward rape. *Sex Roles*, *58*, 790-801.
Ghose, S., & Dou, W. (1998). Interactive functions and their impacts on the appeal of Internet presence sites. *Journal of Advertising Research, 38*(2), 29-43.

Gottschalk, L. A., & Bechtel, R. J. (Eds.). (2008). *Computerized content analysis of speech and verbal texts and its many applications*. New York: Nova Science Publishers.

Gottschalk, L. A., & Bechtel, R. J. (2002). *PCAD 2000: Psychiatric content analysis and diagnosis*. Corona del Mar, CA: GB Software. Retrieved on April 17, 2010 from: http://www.gb-software.com/PCADManual.pdf

Gottschalk, L. A., & Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley, CA: University of California Press.

Grasmuck, S., Martin, J., & Shanyang, Z. (2009). Ethno-Racial Identity Displays on Facebook. *Journal of Computer-Mediated Communication*, *15*(1), 158-188.

Gregory, R. L., with the assistance of Zangwill, O. L. (Eds.). (1987). *The Oxford companion to the mind*. Oxford, UK: Oxford University Press.

Hajek, M., & Kabele, J. (2010). Dual discursive patterns in Czech activists' internet media communication. *European Journal of Communication*, *25*(1), 43-58.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*, 1-23.

Hanna, J. F. (1969). Explanation, prediction, description, and information theory. *Synthese*, *20*, 308-334.

Hanson, N. R. (1959). On the symmetry between explanation and prediction. *The Philosophical Review*, *68*, 349-358.

Hart, R. P. (2000). *The text-analysis program: Diction 5.0*. Austin, TX: Digitext.

Hempel, C. G. (1965). *Aspect of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hu, G., Pan, W., Lu, M., & Wang, J. (2009). The widely shared definition of e-government: An exploratory study. *The Electronic Library*, *27*, 968-985.

Jinsuk, K., Klautke, H., & Serota, K. (2009). Effects of Relational Motivation and Age on Online Self-Disclosure: A Content Analysis of MySpace Profile Pages. *Conference Papers -- International Communication Association*, 1-26. Retrieved on May 1, 2010 from Communication & Mass Media Complete database.

Jones, S., Millermaier, S., Goya-Martinez, M., & Schuler, J. (2008). Whose space is MySpace? A content analysis of MySpace profiles. *First Monday*, *13*(9), 1.

Kane, C. (2008). *I'll see you on MySpace: Self-presentation in a social networking web site*. Unpublished masters thesis, Cleveland State University, Cleveland, OH.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. San Francisco: Chandler Publishing Company.

Kim, I.-H., Anderson, R. C., Nguyen-Jahiel, K., & Archodidou, A. (2007). Discourse patterns during children's collaborative online discussions. *The Journal of the Learning Sciences*, *16*, 333-370.

Kinney, N. T. (2005). Engaging in "loose talk": Analyzing salience in discourse from the formulation of welfare policy. *Policy Sciences*, *38*(4), 251-268.

Kobayashi, J., Spitzberg, B., & Andersen, P. (2008). Communication Predictors of Suicide: The Personification of Suicide in MySpace.com Websites. *Conference Papers -- National Communication Association*, 1. Retrieved on May 1, 2010 from Communication & Mass Media Complete database.

Koertge, N. (1992). Explanation and its problems. *The British Journal for the Philosophy of Science*, *43*(1), 85-98.

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

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management*, *44*, 1448-1466.

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers volume 1*. Cambridge: Cambridge University Press.

Lauzen, M. M., Dozier, D. M., & Cleveland, E. (2006). Genre matters: An examination of women working behind the scenes and on-screen portrayals in reality and scripted prime-time programming. *Sex Roles*, *55*, 445-455.

Lieberman, M. A. (2008). Effects of disease and leader type on moderators in online support groups. *Computers in Human Behavior*, *24*, 2446-2455.

Lieberman, M. A., & Goldstein, B. A. (2006). Not all negative emotions are equal: The role of emotional expression in online support groups for women with breast cancer. *Psycho-Oncology*, *15*, 160-168.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, *52*, 481-495.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed.) (pp. 627-666). Boca Raton: CRC Press.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research*, *30*, 434-437.

Lunk, B. (2008). *MySpace or OurSpace: A cross-cultural empirical analysis of MySpace comments*. Unpublished masters thesis, Cleveland State University, Cleveland, OH.

Matthews, C. (1998). *An introduction to natural language processing through Prolog*. London: Longman.

McAdams, D. P., & Zeldow, P. B. (1993). Construct validity and content analysis. *Journal of Personality Assessment*, *61*(2), 243-245.

McClelland, D. C., Koestner, R., & Weinberger, J. (1992). How do self-attributed and implicit motives differ? In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 49-72). Cambridge: Cambridge University Press.

Mei, W. (2008). Measuring political debate on the Chinese internet forum. *Javnost---The Public, 15*(2), 93-110.

Naccarato, J. L., & Neuendorf, K. A. (1998). Content analysis as a predictive methodology: Recall, readership, and evaluations of business-to-business print advertising. *Journal of Advertising Research, 38*(3), 19-33.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Neuendorf, K. A. (2009). Reliability for content analysis. In A. B. Jordan, D. Kunkel, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis* (pp. 67-87). New York: Routledge.

Neuendorf, K. A. (2011). *The content analysis guidebook* (2nd ed.). Thousand Oaks, CA: Sage.

Neuendorf, K. A., & Skalski, P. D. (2009). Quantitative content analysis and the measurement of collective identity. In R. Adbelal, Y. M. Herrera, A. I. Johnston, & R. McDermott (Eds.), *Measuring identity: A guide for social scientists* (pp. 203-236). Cambridge, MA: Cambridge University Press.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Analysis of affect expressed through the evolving language of online communication. *2007 International Conference on Intelligent User Interfaces, 4564*, 278-281. Retrieved on May 1, 2010 from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.80.5890

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. *Proceedings of the Third International ICWSM Conference* (pp. 278-281). Retrieved on May 1, 2010 from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.4587

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin, 29*, 665-675.

Newton, A. T., Kramer, A. D. I., & McIntosh, D. N. (2009). Autism online: A comparison of word usage in bloggers with and without autism spectrum disorders. *CHI2009: Proceedings of the 27th Annual CHI Conference on Human Factors in Computing Systems, Vols. 1-4*, 463-466.

Norris, P. (2003). Preaching to the converted? Pluralism, participation and party websites. *Party Politics, 9*(1), 21-45.

Opoku, R. A., Pitt, L. F., & Abratt, R. (2007). Positioning in cyberspace: Evaluating bestselling authors' online communicated brand personalities using computer-aided content analysis. *South African Journal of Business Management, 38*(4), 21-32.

Oshita, M. (2010). Generating animation from natural language texts and semantic analysis for motion search and scheduling. *The Visual Computer*, *26*, 339-352.

Pasadeos, Y., Huhman, B., Standley, T., & Wilson, G. (1995, May). *Applications of content analysis in news research: A critical examination*. Paper presented to the Communication Theory and Methodology Division of the Association for Education in Journalism and Mass Communication, Washington, DC.

Patchin, J., & Hinduja, S. (2010). Trends in online social networking: adolescent use of MySpace over time. *New Media & Society*, *12*(2), 197-216.

Pena-Shaff, J. B., & Nicholls, C. (2004). Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers and Education*, *42*, 243-265.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count (LIWC2007)*. Austin, TX: www.licw.net.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*, 258-284.

Radwin, L. E., & Cabral, H. J. (2010). Trust in Nurses Scale: Construct validity and internal reliability evaluation. *Journal of Advanced Nursing*, *66*, 683-689.

Raskin, V. (1985). *Semantic mechanisms of humor*. Dordrecht: Reidel.

Rescher, N. (1997). Hempel-Helmer-Oppenheim, an episode in the history of scientific philosophy in the 20th century. *Philosophy of Science*, *64*, 334-360.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Lawrence Erlbaum.

Rogers, E. M. (1994). *A history of communication study: A biographical approach*. New York: Free Press.

Ruben, D.-H. (1990). *Explaining explanation*. New York: Routledge.

Sagan, C. (1997). *The demon haunted world: Science as a candle in the dark*. New York: Ballantine Books.

Salmon, W. C. (1999). The spirit of logical empiricism: Carl G. Hempel's role in twentieth-century philosophy of science. *Philosophy of Science*, *66*, 333-350.

Shapiro, G., & Markoff, J. (1997). A matter of definition. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 9-34). Mahwah, NJ: Lawrence Erlbaum.

Shoemaker, P. J., & Reese, S. D. (1996). *Mediating the message: Theories of influences on mass media content* (2nd ed.). White Plains, NY: Longman.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods*, *13*, 320-347.

Smith, C. P. (Ed.). (1992). *Motivation and personality Handbook of thematic content analysis*. Cambridge: Cambridge University Press.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communications. *Management Science*, *32*, 1492-1512.

Stepchenkova, S., Kirilenko, A. P., & Morrison, A. M. (2009). Facilitating content analysis in tourism research. *Journal of Travel Research*, *47*, 454-469.

Stone, P. J, Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge: MIT Press.

Tian, Y. & Stewart, C. (2005). Framing the SARS crisis: A computer-assisted text analysis of CNN and BBC online news reports of SARS. *Asian Journal of Communication*, *15*(3), 289-301.

Viegas, F. B., boyd, D., Nguyen, D. H., Potter, J., & Donath, J. (2004). Digital artifacts for remembering and storytelling: *PostHistory* and *Social Network Fragments*. *Proceedings of the 37th Hawaii International Conference on System Sciences* (pp. 1-10). Retrieved on May 1, 2010 from: http://alumni.media.mit.edu/~fviegas/papers/posthistory_snf.pdf

Walther, J. B. (2004). Language and communication technology: Introduction to the special issue. *Journal of Language and Social Psychology*, *23*, 384-396.

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior*, *23*, 2538-2557.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (3rd ed., pp. 529-563). Thousand Oaks, CA: Sage.

Waters, R., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review*, *35*(2), 102-106.

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage Publications.

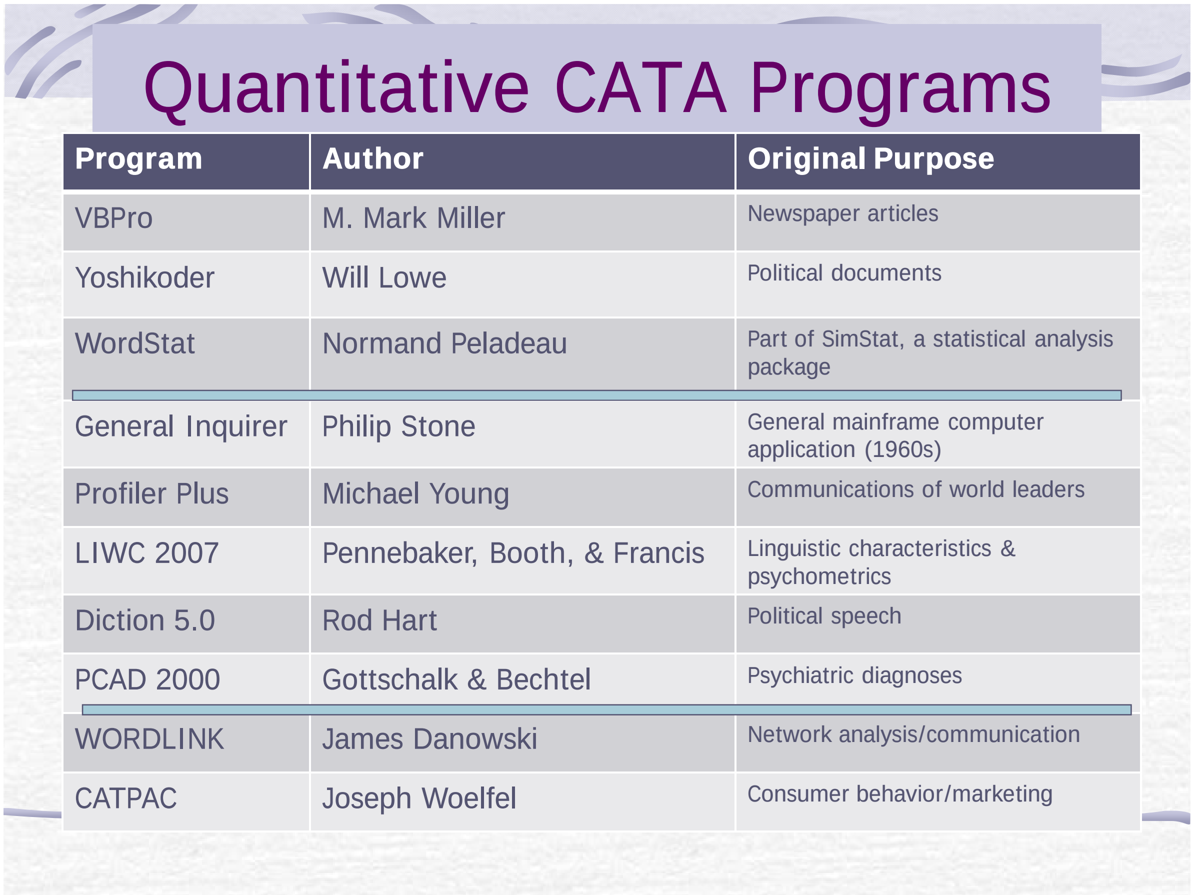
West, M. D. (Ed.). (2001). *Theory, method, and practice in computer content analysis*. Westport, CT: Ablex.

Woelfel, J. (1993). *GalileoCATPAC: User manual and tutorial*. Amherst: Galileo Company. Retrieved on April 17, 2010 from: http://www.galileoco.com/Manuals/CATPAC.PDF

Zhou, L., Burgoon, J. K., Nunamaker, J. F. Jr., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*, 81-106.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F. Jr. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, *20*(4), 139-166.

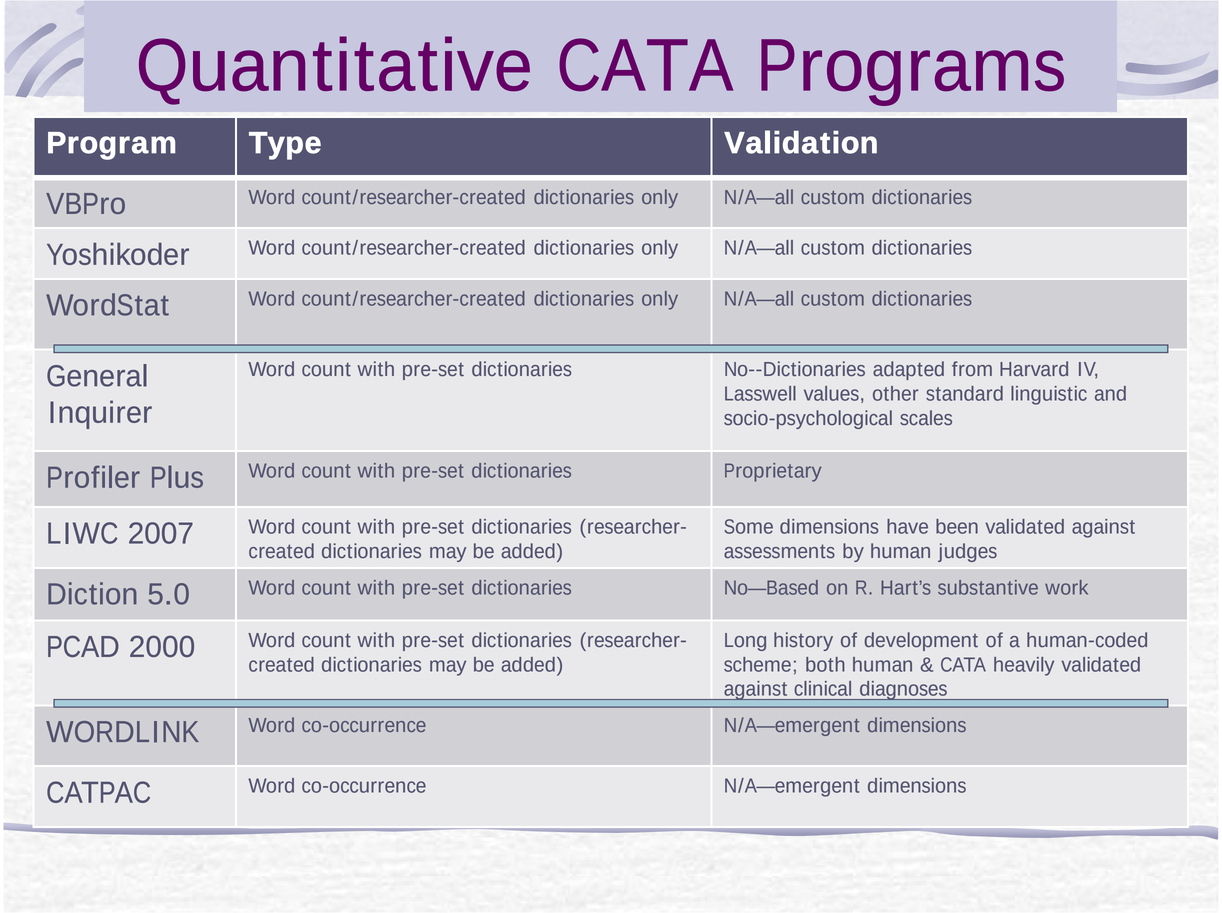
Table 1
Common CATA Computer Programs



Quantitative CATA Programs

Program	Author	Original Purpose
VBPro	M. Mark Miller	Newspaper articles
Yoshikoder	Will Lowe	Political documents
WordStat	Normand Peladeau	Part of SimStat, a statistical analysis package
General Inquirer	Philip Stone	General mainframe computer application (1960s)
Profiler Plus	Michael Young	Communications of world leaders
LIWC 2007	Pennebaker, Booth, & Francis	Linguistic characteristics & psychometrics
Diction 5.0	Rod Hart	Political speech
PCAD 2000	Gottschalk & Bechtel	Psychiatric diagnoses
WORDLINK	James Danowski	Network analysis/communication
CATPAC	Joseph Woelfel	Consumer behavior/marketing

Table 2
CATA Programs: Types and Validation



Quantitative CATA Programs

Program	Type	Validation
VBPro	Word count/researcher-created dictionaries only	N/A—all custom dictionaries
Yoshikoder	Word count/researcher-created dictionaries only	N/A—all custom dictionaries
WordStat	Word count/researcher-created dictionaries only	N/A—all custom dictionaries
General Inquirer	Word count with pre-set dictionaries	No--Dictionaries adapted from Harvard IV, Lasswell values, other standard linguistic and socio-psychological scales
Profiler Plus	Word count with pre-set dictionaries	Proprietary
LIWC 2007	Word count with pre-set dictionaries (researcher-created dictionaries may be added)	Some dimensions have been validated against assessments by human judges
Diction 5.0	Word count with pre-set dictionaries	No—Based on R. Hart's substantive work
PCAD 2000	Word count with pre-set dictionaries (researcher-created dictionaries may be added)	Long history of development of a human-coded scheme; both human & CATA heavily validated against clinical diagnoses
WORDLINK	Word co-occurrence	N/A—emergent dimensions
CATPAC	Word co-occurrence	N/A—emergent dimensions

Endnotes

[i] A list of CATA programs may be found at the web site in support of *The Content Analysis Guidebook* (Neuendorf, 2002; http://academic.csuohio.edu/kneuendorf/content).

[ii] Manifest content may be defined as elements that are present and directly identifiable, while latent content constitutes the deeper meaning, that not directly observable. Based on Freud's interpretation of dreams (Gregory, 1987), the delineation of latent and manifest content is a rather contested approach within content analysis (Potter & Levine-Donnerstein, 1999; Riffe, Lacy, & Fico, 2005; Shapiro & Markoff, 1997). Further, some scholars propose that variables be situated on a continuum (Neuendorf, 2002; Riffe, Lacy, & Fico, 2005) rather than placed in one category or another.

As with surveys and experiments, latent content in content analysis is frequently measured with multiple indicators of manifest characteristics that together represent a latent state (e.g., Radwin & Cabral, 2010), such as Ghose and Dou's (1998) 23 manifest indicators for the latent construct "interactivity" of web sites, and Kinney's (2005) factor analytic extraction of four latent patterns from a set of 11 manifest CATA measures of word use in seven U.S. newspapers.

[iii] The linking of message source data with content analysis message data may allow the discovery of factors important to the process of message generation. An interesting example of this type of study is Lauzen, Dozier, and Cleveland's (2006) investigation of how the involvement of women behind the scenes in the production of reality and scripted prime-time U.S. television programming relates to female representations and portrayals. The presence of women in top creative positions for scripted sitcoms and dramas predicted greater female character representation, and a more egalitarian approach to conflict resolution; these relationships did *not* emerge for reality programming.

The second integrative option, combining content analysis message data with message receiver data, affords an opportunity to test message effects theories. For example, Franiuk, Seelfelt, and Vandello (2008) studied the prevalence of rape myth endorsements in online newspaper headlines about the 2003-2004 Kobe Bryant case, and then conducted an experiment that found male subjects to hold higher rape-supportive attitudes after exposure to myth-endorsing headlines identified via this content analysis.

[iv] However, idiographic approaches are important means of deep understanding of phenomena, and can richly inform more generalized, nomothetic approaches. In an intriguing example of an idiographic application of natural language processing, Taylor and Mazlack (2007) applied a Script-based Semantic Theory of Humor (Raskin, 1985) to individual jokes based on homophonics, aimed at computer recognition of such jokes. However, the ultimate goal of such "computation humor" research is to be able to develop nomothetic principles that will lead to the computational ability to (a) discriminate humorous and non-humorous natural language, (b) provide assistance to children and second-language learners in the mastery of language, (c) facilitate the computer generation of humor, for human-computer interaction and other uses (Binsted

et al., 2006). Importantly, all of these future goals may quite easily have both idiographic and generalized (nomothetic) implications.

[v] Electronic textual discourse may be one-to-one or mass in intent—i.e., messages may be designed for particular receivers, or prepared for a large, undifferentiated audience. This equanimity of interpersonal/mass intent for online messaging is a unique situation in the history of new communication technologies. To date, research has not addressed the differences between online postings intended for targeted receivers and those intended for mass exposure.