International Journal of Artificial Intelligence in Education (2000), 11, to appear

Methodological Issues in the Content Analysis of Computer Conference Transcripts

Liam Rourke, Terry Anderson, D. R. Garrison and Walter Archer *University of Alberta, 3-102 Education North, Edmonton AB, Canada T6G 2G5*
E-mail: lrourke@ualberta.ca

Abstract. This paper discusses the potential and the methodological challenges of analyzing computer conference transcripts using quantitative content analysis. The paper is divided into six sections, which discuss: criteria for content analysis, research designs, types of content, units of analysis, ethical issues, and software to aid analysis. The discussion is supported with a survey of 19 commonly referenced studies published during the last decade. The paper is designed to assist researchers in using content analysis to further the understanding of teaching and learning using computer conferencing.

SCENARIO

Professor Jones has just completed her first university course delivered entirely on-line. The 13-week semester class has left Jones in a state of mild exhaustion. However, the course is finished, the marks have been assigned, and now, thinks Jones, time for some reflection, analysis and perhaps a publishable paper. Jones smiles, confident in the knowledge that the complete transcript of messages exchanged during the course has been captured in machine-readable format. She feels that this accessible data will confirm her hypothesis that students in the on-line course had engaged in much higher levels of discourse and discussion than any she had experienced in ten years of face-to-face instruction. Further, she is interested in investigating the impact of the collaborative learning activity that she instituted in the middle of the course.

Jones is quickly disappointed. The 13-week discussion generated 950 messages. Merely reading them takes her four days. Attempts at cutting and pasting illustrations of higher level thinking into a word processor, have resulted in a hodge-podge of decontextualized quotations, each disparate enough to have Professor Jones questioning her own definitions of higher order thinking. Realizing that the analysis is going nowhere, Professor Jones goes back to the literature and finds a set of criteria laid down by an expert in the field that define the broad areas of thinking skills she sees being developed in the transcripts. Heartened, but now running out of time Professor Jones hires two graduate students to review the messages and identify the incidents of higher order thinking as defined by the expert. Two weeks later, the students report their results: not only have they failed to agree on 70% of the categorizations, but one student has identified 2032 incidents in the transcript, while the other has found only 635 incidents. To add to her misery, Professor Jones also learns that her University's ethics committee, concerned with the large increase in use of computer conferencing for credit courses, has ruled that without informed consent from students, her analysis does not conform with the guidelines of the university's ethical research policy. Feeling overwhelmed and depressed, Professor Jones returns to the educational literature once again, only to find that most of the methodological issues she has been dealing with have not been addressed by major researchers in the field. She also finds that

> *there is no coherent, long-term tradition of researchers who have resolved the methodological problems inherent the analysis of transcripts of text-based computer conferences.*
>
> *This paper is written for the Professor Jones's of the world, hoping that it will help them to release the educational treasures that we believe are locked in the transcripts that document learning in the on-line environment.*

The capacity of computer conferencing to support interaction among participants while providing for temporal and spatial independence creates a unique and valuable environment for distance, distributed, and lifelong learning applications. Additionally, the automatically recorded and machine-readable data generated by this technology offers a compelling source of data for educational researchers and software developers. This paper surveys the efforts of researchers to extract meaning from this data using a research technique called *quantitative content analysis*. Quantitative content analysis is "a research technique for the objective, systematic, quantitative description of the manifest content of communication" (Berelson, 1952, p. 519). Despite the potential of this technique, researchers who have used it have described it as difficult, frustrating, and time-consuming. Very few have published results derived from a second content analysis.

This paper is, therefore, not a meta-analysis of results, but rather an examination of the issues related to the application of this research technique. The intent is to document the evolution of content analysis as it has been used by us and others to analyze transcripts of asynchronous, text-based, computer conferencing in educational settings. These modifiers 'asynchronous,' 'text-based,' and 'educational' provide a definitive focus for our survey. Unfortunately, this prohibits the discussion of some excellent content analysis studies; however, we feel that the use of conferencing in formal education is unique, and that the corpus of studies in this domain is sufficiently large to justify a focal review. We hope that the results of our review and commentary will facilitate the larger goal of improving the quality of teaching and learning through use of this medium.

This paper explores six fundamental issues of content analysis through reference to 19 influential content studies published over the last decade (see Table 1). Section one examines three criteria of quantitative content analysis--objectivity, reliability, and systematic consistency. Section two contrasts the two most common research designs--descriptive and experimental. Section three distinguishes between manifest content and latent content, and section four examines the process of transforming transcripts into unit of data. Sections five and six discuss software packages that facilitate content analysis and ethical issues such as informed consent. The objective of this paper is to provide subsequent researchers with a privileged starting point for their content analysis studies and to contribute to the refinement of this powerful technique.

Table 1. Survey of 19 computer mediated communication content analysis studies

Study	Unit of Analysis	Variables Investigated	Reliability	Research Design
Ahern, Peck, & Laycock (1992)	Message	Interaction Complexity of response	Percent agreement	Descriptive Eperimental
Blanchette (1999)	Thematic	Linguistic variation Participation Discussion themes	Not reported	Descriptive Quasi-experimental
Bullen (1998)	Thematic	Participation Critical thinking	Not reported	Descriptive
Craig et al (2000)	Proposition	Student question type	Percent agreement	Experimental
Fahy et al. (2000)	Sentence	Interaction Participation Critical thinking	Percent agreement	Descriptive
Garrison, Anderson, & Archer (2000b)	Message	Critical Thinking	Cohen's kappa	Descriptive
Hara, Bonk & Angeli (2000)	Paragraph	Participation Interaction Social, cognitive, metacognitive elements	Percent agreement Coder stability	Descriptive
Henri (1991)	Thematic	Participation Interaction Social, cognitive, metacognitive elements	Not reported	Descriptive
Hillman (1999)	Sentence	Patterns of interaction	Cohen's kappa	Descriptive
Howell-Richarson & Mellar (1996)	Illocutionary act	Participation Illocutionary properties Focus (group/task)	Not reported	Descriptive Quasi-experimental
Kanuka & Anderson (1998), Anderson & Kanuka (1997)	Thematic	Collaborative knowledge construction	Not reported	Descriptive
Marttunen (1997, 1998)	Message	Levels of argumentation/ counter argumentation	Reliability coefficient	Descriptive Quasi-experimental
McDonald (1998)	Thematic	Participation Interaction Group development Social, cognitive, metacognitive elements	Cohen's kappa	Descriptive
Mower (1996)	Message	Interaction Topics	[a]Percent agreement after discussion	Descriptive
Newman, Webb, & Cochrane (1995)	Thematic	Critical thinking	Percent agreement after discussion	Descriptive
Rourke et al (in press)	Thematic	Social interaction	Percent agreement	Descriptive
Weiss & Morrison (1998)	Thematic and Message	Critical thinking Understanding/ correcting misunderstandings Emotion	Percent agreement after discussion	Descriptive
Zhu (1997)	Thematic	Interaction Participation Participant roles Knowledge construction	Not reported	Descriptive

Note.

Units of analysis for studies in which participation was described quantitatively are not documented in the table. Routinely, the units of analysis for this measure are number of words, messages, or both.

[a]"Percent agreement after discussion" refers to reliability figures that were obtained through discussion between coders.

CRITERIA OF QUANTITATIVE CONTENT ANALYSIS

Quantitative content analysis can be reduced to four essential steps. Once researchers have a construct they wish to examine, the first step is to identify representative samples of the communication they wish to study. In traditional education studies, this has typically involved making audio or video recordings of classroom interaction between students and teachers, and then transcribing these recordings in preparation for analysis (Flanders, 1970; Sinclair & Coulthard, 1975). In computer mediated communication (CMC) research, this intermediate step is unnecessary because the bulk of current computer conferencing communication is text-based in machine-readable form. Thus, analysis begins with the compilation of selections of transcripts or entire transcripts into text files. The second step involves creating a protocol for identifying and categorizing the target variable(s), and training coders to use this protocol. After a transcript has been coded, the coders' decisions are compared for reliability, and their data is analyzed either to describe the target variable(s), or to identify relationships between variables. The extent to which the resulting descriptions or relationships are valid will depend largely on four criteria discussed in the next subsections--objectivity, reliability, replicability, and systematic coherence.

Objectivity

Berelson (1952) stipulates that content analysis is an objective technique. In the context of content analysis, *objective* refers to the extent to which categorization of sections of transcripts is subject to influence by the coders. This technique, perhaps more so than any other quantitative technique, is susceptible to the infiltration of subjectivity and interpretive bias. Mower's (1996) candid discussion of reliability is illustrative:

> In instances of disagreement, [rater 1] agreed that [rater 2's] evaluation could be correct. In other instances of disagreement, it was determined that remarks could fit into either one of two categories depending upon the [rater's] interpretation. It was concluded that sometimes, subjective judgment was involved in assigning some topics to categories" (p. 220).

Mower's frankness reveals a pervasive issue in content analysis studies. While some amount of subjectivity may be unavoidable in coding transcripts, a quantitative study should not conclude with an admission that objectivity and reliability have not been achieved. Rather, the discovery of an excessive degree of subjectivity should signal to the research team that further refinement is needed in category definition or coding protocol.

Reliability

The primary test of objectivity in content studies is *interrater reliability*, defined as the extent to which different coders, each coding the same content, come to the same coding decisions. Potter & Levine-Donnerstein (1999) regard reliability data as an important part of content reports and offer the following advice: "If content analysts cannot demonstrate strong reliability for their findings, then people who want to apply these findings should be wary of developing implementations" (p. 258). Of the 19 published studies in our sample, only ten reported reliability data (see Table 1).

The simplest and most common method of reporting interrater reliability is the percent agreement statistic. This statistic reflects the number of agreements per total number of coding decisions. Holsti's (1969) coefficient of reliability (C. R.) provides a formula for calculating percent agreement:

$$C.\,R. = 2m / n_1 + n_2$$

Where: m = the number of coding decisions upon which the two coders agree

n_1 = number of coding decisions made by rater 1

n_2 = number of coding decisions made by rater 2

Many statisticians characterize interjudge agreement as inadequate because it does not account for chance agreement among raters (Capozzoli, McSweeney, & Sinha, 1999). Three of the studies in our sample used the *Cohen's kappa* (k) statistic to determine reliability. Cohen's kappa is a chance-corrected measure of interrater reliability that assumes two raters, *n* cases, and *m* mutually exclusive and exhaustive nominal categories (Capozzoli, McSweeney, & Sinha). The formula for calculating kappa is:

$k = (F_o - F_c) / (N - F_c)$
Where: N = the total number of judgements made by each coder
F_o = the number of judgements on which the coders agree
F_c = the number of judgements for which agreement is expected by chance.

In Cohen's (1960) original formula, agreement by chance is calculated in four steps. Researchers begin by counting the number of times a category of a coding scheme is used by the coders. Then, this figure is converted to a percentage of all coding decisions. Finally, this percentage is squared, and the squared percentages for all categoies are summed (see Capozzoli, McSweeney, & Sinha, 1999; Cohen, 1960; and Potter & Levine-Donnerstein, 1999 for further discussion).

Although kappa is a powerful measure of interrater reliability, some authors have argued that it is overly conservative (Potter & Levine-Donnerstein, 1999). This is particularly true with coding protocols that include several categories, thereby making the possibility of chance agreement negligible. Further, as Hagelin (1999) suggests, "factors such as the number of observations, the number of categories, and the distribution of the data influence the kappa values in such a way as to make interrater agreement difficult to interpret" (p. 314).

The exact level of interrater reliability that must be achieved has not been clearly established. For Cohen's kappa, Capozzoli, McSweeney, and Sinha (1999) declare that:

> …values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance (p. 6).

For percent agreement figures, Riffe, Lacy, and Fico (1998) state that, in communication research, "a minimum level of 80% is usually the standard" (p. 128). Beyond the statistical context, content analysts suggest that researchers must decide for themselves the level of acceptable agreement. Riffe, Lacy, and Fico add the following: "Research that is breaking new ground with concepts that are rich in analytical value may go ahead with reliability levels somewhat below that range" (p. 131). This lenience is based on the premise that some measures that are taken to increase reliability may simultaneously reduce the value of the results, or in Krippendorf's words "reliability often gets in the way of validity" (1980, p. 130). Garrison, Anderson, and Archer (2000b), Hillman (1999), and McDonald (1998) reported kappa's of 0. 74, 0.96, and 0.67 respectively; however, it is premature to declare a conventional level of acceptability. We feel that the mere act of reporting these figures gives readers sufficient information to interpret results.

The difficulty of achieving acceptable levels of interrater reliability has lead to the development of an alternative system of coding transcripts. Barros and Verdejo (2000), Duffy, Dueber, and Hawley (1998), and Ravenscroft and Pilkington (2000) have developed semi-structured computer conferencing systems, in which participants chose the type of contribution that they are making from a limited set of alternatives. For example, in the issue-based discussion forum of Duffy, Dueber, and Hawley's system, students post a message by selecting one of four labels--Hypothesis, Important Point, Evidence, or Learning Issue. To the content analyst, this means that students are essentially coding their own messages. Barros and Verdejo have developed a parallel system that also includes automatic message analysis features.

Replicability

The reliability of a coding scheme can be viewed as a continuum, beginning with coder stability (one coder agreeing with herself over time), to interrater reliability (two or more coders agreeing with each other), and ultimately to replicability (the ability of multiple and distinct groups of researchers to apply a coding scheme reliably). Thus, the definitive test of a coding scheme is replicability. The coding scheme that has been the basis for most attempts at replication originates in Henri's (1991) seminal article. However, each time her protocol is used, it is criticized and either modified or abandoned (Bullen, 1998; Gunawardena, Anderson, & Lowe, 1997; Hara, Bonk & Angeli, 2000; Howell-Richardson & Mellar, 1996; Kanuka & Anderson, 1998; Newman, Webb, & Cochrane, 1995). The fact that Henri's procedure has drawn criticism is, paradoxically, complimentary; since most researchers explicitly build upon the ideas enunciated by Henri, only her ideas have been criticized. No other model has generated attempts at replication; therefore, no other model has drawn significant criticism. This lack of replication (i.e., of successful applications of other researchers' coding schemes) should be regarded as a serious problem. In most fields, even when a valid study yields statistically significant findings, the results are described cautiously as "supportive" until they have been replicated. Reliable application of a coding scheme by researchers who are not involved in its creation would be a convincing testament to its efficacy. Newman, Webb, and Cochrane (1995) conclude their study with an invitation to other researchers to apply and improve upon their protocol; likewise, Howell-Richardson and Mellar (1996) suggest that the validity of their method "...is an empirical question" (p. 53). With these statements, they are inviting others to test their method in practice. Unfortunately, to our knowledge no one has done so.

Systematic

In the context of content analysis, the term *systematic* refers to "a more or less well structured set of ideas, assumptions, concepts and interpretative tendencies, which serves to structure the data of an area" (Reber, 1995, p. 780). Barros and Verdejo (2000) and Kanuka and Anderson's (1998) studies provide good models of systematic studies. At the outset of their investigation, Kanuka and Anderson recognized an association between the attributes of computer conferencing and the tenets of constructivism. Therefore, they began by identifying their perspective as constructivist and then selected a transcript analysis instrument that views communicative behavior in terms of active, collaborative, construction of knowledge. Other studies, however, combine irreconcilable paradigms in their analysis of data. Howell-Richardson and Mellar (1996) identify this issue in Henri's (1991) classification schema, noting that "the level of description at the social, cognitive skills or interactivity levels was dependent on a mixture of theoretical approaches, which were not necessarily mutually consistent" (p. 69). As late as 1998, Bullen was casting a wide net, alternately sampling from Ennis' (1987) cognitive perspective of critical thinking, Henri's (1991) behaviorist perspective of interactivity, and Harasim's (1990) constructivist perspective of participation. The exploratory nature of these studies reflects the immaturity of the field rather than the deficiencies of the methodology or the nescience of the researchers. As in any new field of research, many of the studies in the sample were descriptive, with only a few experimental efforts. Both of these designs are described in the next section.

RESEARCH DESIGN

Descriptive

Berelson (1952) characterizes content analysis as primarily a descriptive technique. Of the 19 studies we reviewed, 18 were either partially or entirely descriptive - i.e., they described,

organized and summarized what was occurring in a specific computer conference (see Table 1). In these studies, information has been collected on several important themes associated with educational uses of computer conferencing, which gives subsequent researchers a foundation upon which they can build. For example, Bullen (1998) characterized participation in his group as "low to medium" relative to participation levels in Harasim's (1990) study. These studies provide a rich source of anecdotal data and a model for the acquisition of fundamental information.

Experimental

Often researchers want to extend the purpose of content analysis from simple description to inferential hypothesis testing. Borg and Gall (1989) discuss this shift in the context of educational research: "Whereas most early studies employing content analysis relied on simple frequency counts of objective variables (e.g., spelling errors), recent studies more often aim at using content analysis to gain insights into complex social and psychological variables" (p. 521). To this, they add the following caveat: "Such studies are much more difficult to carry out than the simple frequency studies and often depend on a researcher's high level of sophistication" (p. 521).

Ahern, Peck, and Laycock's study of 1992 was the first in this domain to combine the content analysis technique with random assignment to groups and controlled manipulation of variables. This approach was advanced by Craig, Gholson, Ventura, and Graesser (2000), Howell-Richardson and Mellar (1996), and Marttunen (1997, 1998) and who were able to draw convincing conclusions concerning different experimental or quasi-experimental conditions.

Our discussion now turns to the selection of the object of investigation. There is much to be learned from the study of both the manifest of latent content of transcripts; however, each of these types of content presents measurement challenges.

NATURE OF CONTENT

Manifest content

Manifest content is content that resides on the surface of communication and is therefore easily observable. An example of analyzing manifest content is provided by Rourke et al (in press) who counted the number of times students addressed each other by name. The coding of manifest content can (at least in theory) be sufficiently formalized so as to be undertaken by machines, and imposes little interpretive burden upon coders (Hagelin, 1999). This ease of coding makes it attractive for content analysis. Berelson (1952), Holsti (1969), and Riffe, Lacy, and Fico (1998) concur that "the requirements of scientific objectivity dictate that coding be restricted to manifest content" (Holsti, p. 12). Several important conferencing issues have been studied in this manner including participation, interaction, use of emoticons, and linguistic variation (see Table 1). Doubtless, there are other manifest behaviors of interest to scholars of computer conferencing interactions that will be measured and described in future studies.

Latent Content

Not all research questions, and especially not many of the most interesting ones, can be answered by focusing on the manifest or surface content of the transcripts. The overriding concern of many educational researchers is whether or not computer conferencing can facilitate higher-order learning outcomes, which educational theorists are coming to regard not as overt products, but rather as covert processes (Anderson & Garrison, 1995).

As early as 1951, Bales struggled with the problems of measuring latent behavior. In his study of face-to-face groups, Bales used two coders with a third operating an audio recorder for reliability checking. Coders using his system had to code in real time and were less able to, and

less interested in, the analysis of long passages or sequential series of interactions. As Bales notes "in a sense, the coder must work more or less on the surface meaning of activity, and forgo involved in-depth interpretations" (p. 35). Nonetheless, Bales rejected mechanist counting of manifest variables in interaction analysis. He sought an interpretive analysis of behavior that "involves the imputation of meaning, 'the reading in' of content, the inference that the behavior has function(s) either by intent or effect" (p. 6). Like Bales, many educational researchers, including ourselves, are more interested in struggling with the important (though hidden) facets of individual and social cognition rather than assessing that which is most easily measured. Fortunately the temporal constraints and the necessity of staying attuned to non-verbal interaction that coders struggled with using Bale's system have been eliminated in the analysis of CMC transcripts.

Potter and Levine-Donnerstein (1999) make a further distinction between two types of latent content. The first they call *latent pattern* variables. As an example, they offer "style of dress: formal or casual." In making this type of coding decision, coders resort to an inventory of clues (e.g. ties, jewelry, etc.) that indicate the possible existence of the target variable; however, this possibility is confirmed only when other elements or an appropriate pattern of elements is concurrently present. An example of latent pattern variable analysis is provided by Marttunen (1997, 1998), who studied argumentation and counter-argumentation in students' email messages. Arguments were conceptualized as having four properties--claims, grounds, warrant, and rebuttals. The presence of one of these characteristics served to sensitize coders to the possibility that the message could be coded as an argument; however, judgement was withheld until more of the four properties, in an appropriate pattern, were identified.

Coding schemes for latent pattern variables are similar to, but more sophisticated than coding schemes for manifest variables. Interrater reliability increases as the list of indicators and cues approaches completeness, and coders are alert and well trained.

The same principles do not apply to *latent projective* variables. In latent projective variables, the locus of the variable shifts to the coders interpretations of the meaning of the content. This is in contrast to both manifest and latent pattern variables, in which the meaning of the communication, or the target variable, resides on the surface of the content. Rourke et al (in press) included the category "use of humor" in their coding scheme and found that reliable coding depended on the intersubjectivity of coders' social and cognitive schemata. In other words, coders from different cultural backgrounds, ages, and personality types seem to have difficulty in reliably identifying humor and other latent projective variables.

In the studies that we reviewed, cognitive processes were the most commonly investigated latent variable. Henri's (1991) and Zhu's (1996) classification schemata look for "cognitive' dimensions" in the transcripts. Others, beginning with Mason (1991), look for evidence of "critical thinking" as it is variously defined (Bullen, 1998; Fahy et al., 2000; Garrison, Anderson, & Archer, 2000b; and Newman, Web, & Cochrane, 1997).

Experienced content analysts argue that measuring latent content is inherently subjective and interpretative. Henri's taxonomy has been criticized on these grounds by Hara, Bonk and Angeli (1998), Howell-Richardson and Mellar (1996), and Newman, Webb, and Cochrane (1995). Newman, Webb, and Cochrane's coding protocol that accompanies their instrument illustrates clearly the practical problems of identifying latent variables:

> Rather than classify every statement in a transcript as, e.g. critical assessment or uncritical acceptance, we mark and count the obvious examples, and ignore the intermediate shades of grey. This eases the task of the assessors, since there is less need for subtle, subjective, borderline judgements...Of course, one statement might show more than one indicator, ... Or indicators can even overlap (p. 69).

The implications of this protocol on objectivity and reliability are obvious.

Instead of identifying latent variables during coding, Holsti (1969) suggests postponing this type of analysis to the interpretive stage, "at which time," he adds, "the investigator is free to use all of his powers of imagination and intuition to draw meaningful conclusions from the data (pp. 12-13). Two studies have taken this approach. To begin her study, Mason (1991)

induced a typology of common communicative behaviors in conferencing transcripts. Her typology included six items such as "use of personal experiences related to course themes, and reference to appropriate material outside the course package" (p. 168). Mason and Weiss and Morrison (1998) used these manifest elements to code the transcripts. Then, in the final stages of their studies, they proposed an association between the manifest behaviors and latent variables such as critical thinking, judgement, and initiative.

A more popular alternative has been to define the latent variables and then deduce manifest indicators of these variables (Bullen, 1998; Garrison, Anderson, & Archer, 2000b; Gunawardena, Lowe, & Anderson, 1997; Henri, 1991; Marttunen, 1997, McDonald, 1998; 1998; Newman, Webb, & Cochrane, 1995; Zhu, 1996; Rourke et al, in press). For example, Henri's "surface processing" category was identified in the transcript through indicators such as "repeating what has been said without adding any new elements" (p. 130). Both of these approaches, inductive and deductive, have been useful for studying latent variables through a survey of manifest content. Identifying the target variables as manifest or latent will influence the determination of the unit of analysis, a process that is discussed in the next section.

UNIT OF ANALYSIS

Part of conducting a quantitative study involves identifying the segments of the transcript that will be recorded and categorized. In content analysis nomenclature, this process is called *unitizing*. Researchers have experimented with different types of recording units with varying degrees of success. Their goal has been to select a unit that multiple coders can identify reliably, and simultaneously, one that exhaustively and exclusively encompasses the sought-after construct. The research that we reviewed in this article points to a frustrating negative correlation between these two criteria. Fixed units such as single words or entire messages are objectively recognizable, but they do not always properly encompass the construct under investigation. Dynamic units such as Henri's (1991) "unit of meaning" properly delimit the construct, but invite subjective and inconsistent identification of the unit.

Sentence Unit

Units such as the word, proposition, or the sentence are called *syntactical units* because they are delimited by syntactical criteria. Fahy et al. (2000) and Hillman (1999) used the sentence as their recording unit to help meet the goal of developing instruments that are easy to use and reliable. During a preliminary analysis, Fahy et al. reported percent agreement figures as high as 94% and Hillman reported a kappa of 0.96 for the three variables investigated in his study. Our experience with this unit of analysis was less encouraging. The objectivity of a syntactical unit is confounded by the idiosyncratic nature of conferencing communications (Blanchette, 1999). The syntax in the conferences we studied combined the telegraphic style of email with the informality of oral conversation. The following selection from one of our transcripts is typical:

> Certain subjects could be called training subjects...i.e. How to apply artificial respiration....as in first aid...and though you may want to be a guide on the side....one must know the correct procedures in order to teach competency...other subjects lead themselves very well to exploration and comment/research [*ellipses in original*].

How many sentences are present in the preceding transcript selection? The strength of the sentence unit – reliable identification – did not materialize in this example. Use of the sentence unit also introduces an additional subjective step to the research process in that coders must first interpret the messages posted by the participants in the conference and transform them into sentences. Also, sentence level coding yields an enormous number of cases. Hillman reports an average of 8,680 sentences in each of the transcripts he analyzed.

Paragraph Unit

Hara, Bonk, and Angeli (2000) attempted to use a slightly larger syntactical unit, the paragraph. Use of this unit could significantly reduce the number of cases, as compared to the number generated by the use of the sentence unit. However, as the size of the unit expands, so does the likelihood that the unit will encompass multiple variables. Conversely, one variable may span multiple paragraphs. Also, our experience did not support the authors' optimistic statement that "college-level students should be able to break down the messages into paragraphs" (p. 9). Often, a full line of space or a tab was used for purposes other than delimiting a single coherent and unified idea accompanied by a group of supporting sentences. And, once the syntactical criteria are lost, the definition of the unit as "paragraph" becomes meaningless: What the coders are identifying are, in fact, graphical blocks of text. Hara, Bonk and Angeli's (2000) ad hoc coding protocol reveals these problems: "...when two continuous paragraphs dealt with the same ideas, they were each counted as a separate unit. And when one paragraph contained two ideas, it was counted as a two separate units" (p. 9). Using this protocol, Hara, Bonk & Angeli settled for an aggregate percent agreement figure of 74.6%, which was "…deemed adequate given the subjectiveness of such scoring criteria" (p. 9).

Message Unit

Marttunen (1997, 1998) looked for levels of argumentation and counterargumentation in transcripts, and like Ahern, Peck, and Laycock (1992) and Garrison, Anderson, and Archer (2000b) used the message as the unit of analysis. This unit has important advantages. First, it is objectively identifiable: Unlike other units of analysis, multiple raters can agree consistently on the total number of cases. Second, it produces a manageable set of cases. Marttunen and Ahern, Peck, and Laycock recorded a total of 545 and 185 messages respectively, a total that would obviously have been considerably larger if the messages had been subdivided. Third, it exhaustively and exclusively contained the object of Marttunen's and Ahern, Peck, and Laycock's studies. Fourth, it is a unit whose parameters are determined by the author of the message. In the discussion of interrater reliability, Marttunen reported a reliability *(r)* of 0.71; Ahern, Peck, and Laycock reported percent agreement at "over 90%" (p. 298), and Garrison, Anderson, and Archer reported a kappa of 0.74 when using the message unit as the unit of analysis..

Thematic Unit

Unfortunately, the message is not suitable for all variables. The most commonly used unit in our sample was introduced by Henri (1991), who rejected the process of a priori and authoritatively fixing the size of the unit based on criteria that are tangential to the construct under study. Instead, she used a "unit of meaning," which is similar in form to the conventional thematic unit described by Budd, Thorp and Donohue (1967) as "…a single thought unit or idea unit that conveys a single item of information extracted from a segment of content" (p. 34). Henri justifies this approach by arguing that "it is absolutely useless to wonder if it is the word, the proposition, the sentence or the paragraph which is the proper unit of meaning, for the unit of meaning is lodged in meaning (p. 134). However, coding a complex, latent construct such as "in-depth processing" with a volatile unit such the "meaning unit" creates extensive opportunity for subjective ratings and low reliability. Not surprisingly, Henri offers no reliability discussion.

Illocutionary Unit

Howell-Richardson and Mellar (1996) attempted to improve the reliability of this procedure by establishing a theoretical basis for Henri's meaning unit. Drawing on Speech-Act theory, Howell-Richardson and Mellar explained that transcripts should be viewed with the following

question in mind: What is the purpose of a particular utterance? A change in purpose sets the parameters for the unit. The authors evaded some of the difficulties that Henri's scheme presents by focusing on manifest content such as the linguistic properties of a message and the audience to whom it was directed. Howell-Richardson and Mellar's method has advantages; however, rather than reporting interrater reliability figures, the authors submit the following tantalizing discussion:

> Our procedure overcomes both the problem of relying on potentially inconsistent judgements in deciding whether or not a set of wordings constitute a single meaning or more than one and the problem of suggesting that graphic boundaries of the message can be equated with a single communicative act (p. 52).

The selection of the unit of analysis is complex and challenging for the quantitative content analysis researcher. Krippendorf (1980) concedes that, ultimately, the process of unitization "involves considerable compromise" (p. 64) between meaningfulness, productivity, efficiency, and reliability.

SOFTWARE TO AID CONTENT ANALYSIS

The existence of machine-readable data (the conference transcripts) does not guarantee that the transcripts are available in a format that can easily be analyzed. A first problem is gathering the data into a single text file that contains the entire sampling unit. Some conferencing software does not support export of the complete conference or selected portions, but rather forces researchers to tediously cut and paste each individual message from a separate window into a larger text file.

Once the data have been moved into a text file, there are a number of software packages that can be used to assist in the process of analysis. The most useful are qualitative analysis packages such as Atlas/ti®, NUD*IST® and HyperQual®. These packages allow the researcher to identify the unit of analysis in the transcript and assign the text to a coding category that has been theoretically defined apriori or to one that emerges from the analysis process. Later analysis can combine or sort codes into families for more meaningful discussion, presentation or analysis. These packages allow multiple coding of individual passages for use when more than one construct is being investigated and also allow multiple coders to work on a single coding task while maintaining identification of the coder for calculation of reliability. A wide variety of reports can be generated from these packages including list and frequency counts of codes with or without illustrative quotations from the text.

In addition to the hand coding by researchers, many of these packages allow coding to be automated, based upon multi-string text search and pattern matching. Other quantitative data can be generated, including number of sentences, coding results by individual posters, and counts of results from multiple documents.

Once the content of the transcripts had been coded and categorized, SPSS or other statistical programs can be employed for more quantitative analyses and calculation of reliability. Percent agreement calculations are performed using SPSS's chi square function, and final interrater reliability figures can be calculated with SPSS's Cohen's kappa statistic.

ETHICS

We conclude with a brief discussion of the ethical issues related to content analysis of computer conferencing transcripts. Questions of ethical approval and informed consent are important to all researchers and their subjects. We have had personal experience in which a proposed study was funded and then aborted due to the reluctance of a single individual to allow external researchers to review the contents of the computer conference transcript.

Alternatively, we have been involved in the tedious process of obtaining ethical clearance from a university ethics approval board and have been left wondering if the approval was either useful or necessary.

Our experience as researchers in a Canadian university operating under ethical approval guidelines set by our university and required by Canadian federal research granting councils is probably similar to that of researchers operating under other jurisdictions. However, each research team should investigate policies and practices that apply in their particular circumstances.

Ethical guidelines have been established to protect human subjects from harm as a result of participation in scientific investigation. The three Canadian federal granting councils released a Code of Ethical Conduct for Research Involving Humans in 1994 (http://www.mrc.gc.ca/ethics/code/english/toc.html). This Code cites four principles to guide researchers in construction and evaluation of research protocols. These principles are 1) respect for persons, 2) non-maleficence, 3) beneficence, and 4) justice.

The respect for persons principle is grounded upon the right of participants to make informed choice as to degree (if any) of participation in the study. This is the area of greatest issue to many researchers. This code defines research participants as "living individuals or groups of living individuals about whom a scholar conducting research obtains (1) data through intervention or interaction with the individual or group, or (2) identifiable private information". Distinguishing between active action research in which the researcher takes part in the conference under investigation and research projects in which the researcher merely examines the subsequent transcript, changes the nature of the "intervention or interaction" between researcher and research subject. We argue that a researcher analyzing the transcripts of a conference, without participating in the conference, has not intervened in the process and thus has not placed them in the position of research participants. However, the second criterion is relevant in that often transcripts contain "private information" that has been posted to the conferencing group.

Two solutions to this problem are possible. The researcher can request that each participant sign a conventional informed consent release form in which the standard information is provided to participants describing: nature of the investigation, potential harm and benefits, how the information obtained is to be used, and how the participants can contact the researchers to discuss any concerns they may have. This standard process of subject permission is complicated in a formal education context in which protection of privacy may preclude the release of addresses of students to which the researcher can post release forms. In our experience, transmitting such forms by email or posting messages within an administration section of the computer conference results in the majority of students responding positively to the request, none objecting, but a few not replying at all. In the worst case, a negative response, or lack of any response, forces researchers to either abandon this sample group or have the postings of individuals who have not given permission removed from the transcript prior to analysis. Removal of individual postings is possible using search and delete techniques of the analysis software, but in practice becomes problematic in that postings often contain excerpts and quotations from previous postings, any of which may have been made by non-participating subjects. In addition, use of personal names is common and eradicating all references to non-participants can be very time consuming. Further, one could narrowly define removal of a non-participant’s posting itself as an analysis process requiring permission of the participants. Finally, the removal of one or more person's postings may make understanding of the conference text impossible and decontextualize subsequent postings.

A second more encompassing solution is to reduce the requirement for informed consent by applying the two criteria of the "research participant" above and concluding that transcript analysis participants are not, by definition research participants. To arrive at such a conclusion one must address the second stipulation that the researcher not obtain "identifiable private information". The use of "search and replace" features of analysis software is then used to change all personal or login names from headers of postings and within the postings to "subject 1, subject 2" etc.

The study of computer conferencing transcripts seems to present little danger of maleficence, and we believe high potential for beneficence -- especially in potential to increase learning efficacy of subsequent conferences. The issue of justice seems not to be of major concern and is normally an issue only when conducting research with specialized target groups based on gender, race or social economic status. Thus the issue of informed consent seems to be the most problematic ethics issue for transcript analysis researchers. There seems to be no easy solution to this problem, other than for researchers to expect to expend some considerable energy obtaining consent or stripping non-participant postings or personal identification from the transcripts.

CONCLUSION

In 1996, Mason and Romiszowski remarked that:

> The most glaring omission in CMC research continues to be the lack of analytical techniques applied to the content of the conference transcript. Given that the educational value of computer conferencing is much touted by enthusiasts, it is remarkable that so few evaluators are willing to tackle this research area. (p. 443).

As conferencing matures and diffuses, naïve enthusiasm is giving way to practical questions about how this technology can be used to facilitate specific educational objectives. This attitude is leading to a shift in the literature away from anecdotal, promotional essays toward more objective research. We hope that a portion of this research will exploit the quantitative content analysis technique.

The 19 studies that we surveyed have demonstrated the value of this research technique for both descriptive and experimental purposes. The authors of these studies have also addressed many of the methodological problems that hinder the application of this technique to educational computer conferencing transcripts.

The main shortcoming of the quantitative content analysis studies in our sample was the failure of researchers to adhere to the principles that make quantitative research valid. Characteristics such as objectivity and reliability are not accidental features of some studies; rather, they are important criteria for any studies using this technique. As Riffe, Lacy, and Fico (1998) insist "failure to report reliability virtually invalidates whatever usefulness a content study may have" (p. 134).

In our own studies we are examining the use of computer conferencing to support higher order learning through peer and instructor interaction. (Garrison, Anderson & Archer, 2000a; Garrison, Anderson, & Archer, 2000b; Rourke et al., in press). We are attempting to develop transcript analysis tools that are efficient, reliable, valid, and practical so that they may be used to evaluate conference activity not only by researchers, but also by teachers and instructional designers. The task of developing instruments and techniques for transcript analysis that meet these criteria is a necessary prerequisite to the empirical investigation of asynchronous, text-based computer conferencing. Further studies are needed to identify the salient elements this medium. Not all of the original hyperbolic claims for the benefits of computer conferencing have been empirically tested. Does asynchronous communication really foster more reflective and careful response composition? Does text-based communication actually lead to more articulate presentation of arguments? If these claims are supported, then experimental designs will play an important role in defining exactly how to facilitate this potential. To answer these increasingly important questions being asked about the use of computer conferencing in higher education, we need to undertake rigorous and systematic research studies.

Acknowledgements

This study is supported in part by a grant from the Social Sciences and Humanities Research Council of Canada.

References

Anderson, T. & Garrison, D. R. (1995). Critical thinking in distance education: Developing critical communities in an audio teleconference context. *Higher Education*, 29, *183-199.*

Anderson, T. & Kanuka, H. (1997). On-line forums: New platforms for professional development and group collaboration. (ERIC Document Reproduction Service, ED 418 693).

Ahern, T., Peck, K. & Laycock, M. (1992). The effects of teacher discourse in computer-mediated discussion. *Journal of Educational Computing Research, 8*(3) 291-309.

Bales, R. (1951) Interaction process analysis. Cambridge: Addison -Wesley.

Barros, B. & Verdejo, F. (2000). Analyzing student interaction processes in order to improve collaboration: The DEGREE approach. *International Journal of Artificial Intelligence in Education, 11,* to appear.

Berelson, B. (1952). *Content analysis in communication research*. Illinois: Free Press.

Blanchette, J. (1999). Register choice: Linguistic variation in an online classroom. *International Journal of Educational Telecommunications, 5* (2), 127-142.

Borg, W. & Gall, M. (1989). The methods and tools of observational research. In W. Borg & M. Gall (Eds.) *Educational research: An introduction (5th ed.)* (pp. 473-530). London: Longman.

Budd, R., Thorp, R., & Donohew, L. (1967). *Content analysis of communications.* London: The Collier-McMillan Limited.

Bullen, M. (1998). Participation and critical thinking in online university distance education. *Journal of Distance Education, 13(2),* 1-32.

Capozzoli, M., McSweeney, L. & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*(1), 3-23.

Cohen. J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37-46.

Craig, S., Gholson, B., Ventura, M., Graesser, A., & The Tutoring Research Group (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* to appear.

Duffy, T., Dueber, B., & Hawley, C. (1998). Critical thinking in a distributed environment: A pedagogical base for the design of conferencing systems. In C. Bonk & K. King (Eds.) *Electronic collaborators: Learner-centered technologies for literacy, apprenticeship, and discourse (pp. 51-78)*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Ennis, R. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp.9-26). New York: Freeman.

Fahy, P. J., Crawford, G., Ally, M., Cookson, P., Keller, V. and Prosser, F. (2000). The development and testing of a tool for analysis of computer mediated conferencing transcripts. *Alberta Journal of Educational Research, 46*(1), *Spring,* 85-88.

Flanders, N. (1970). *Analyzing teacher behavior*. Reading, MA: Addison-Wesley.

Garrison, D. R., Anderson, T., & Archer, W. (2000a). Critical thinking in a text-based environment. Computer conferencing in higher education. *Internet in Higher Education, 2*(2), 87-105.

Garrison, D. R., Anderson, T., & Archer, W. (2000b). Critical thinking and computer conferencing: A model and tool to assess cognitive presence. Submitted for publication

Gunawardena, C., Lowe, C. & Anderson, T. (1997). Analysis of a global on-line debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 395-429.

Hagelin, E. (1999). Coding data from child health records: The relationship between interrater agreement and interpretive burden. *Journal of Pediatric Nursing, 14*(5), 313-321.

Hara, N., Bonk, C., & Angeli, C., (2000). Content analyses of on-line discussion in an applied educational psychology course. *Instructional Science. 28*(2), 115-152.

Harasim, L. (1990). On-line education: Perspectives on a new environment. New York: Praeger.

Henri, F. (1991). Computer conferencing and content analysis. In A. Kaye (Ed.) *Collaborative learning through computer conferencing: The Najaden papers,* (pp.117-136). London: Springer-Verlag,

Hillman, D. (1999). A new method for analyzing patterns of interaction. *The American Journal of Distance Education, 13*(2), 37-47.

Holsti, O. (1969). *Content analysis for the social sciences and humanities*. Don Mills: Addison-Wesley Publishing Company.

Howell-Richardson, C. & Mellar, H. (1996). A methodology for the analysis of patterns of participation within computer mediated communication courses. *Instructional Science, 24*, 47-69.

Kanuka, H. & Anderson, T. (1998). Online social interchange, discord, and knowledge construction. *Journal of Distance Education, 13*(1) 57-75.

Krippendorf, K. (1980*). Quantitative content analysis: An introduction to its method*. Beverly Hills: Sage Publications.

Marttunen, M. (1998). Learning of argumentation in face-to-face and e-mail environments. (ERIC Document Reproduction Service, ED 422 791).

Marttunen, M. (1997). Electronic mail as a pedagogical delivery system. *Research in Higher Education, 38(3)*, 345-363.

Mason, R. (1991). Analyzing computer conferencing interactions. *Computers in Adult Education and Training, 2*(3), 161-173.

Mason, R., & Romiskowski, A. (1996). Computer-mediated communication. In D. Jonassen (Ed.). *Handbook of Research for Educational Communications and Technology*. New York: Macmillan.

McDonald, J. (1998). Interpersonal group dynamics and development in computer conferencing: The rest of the story. In *Wisconsin Distance Education Proceedings* Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.
Available [Online]: http://www.mrc.gc.ca/publications/publications.html

Mower, D. (1996). A content analysis of student/instructor communication via computer conferencing. *Higher Education, 32,* 217-241.

Newman, G., Johnson, C., Webb, B. & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co-operative learning. *Journal of the American Society for Information Science, 48(6),* 484-495.

Newman, G., Webb, B. & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology, 3*(2), *56-77.*
Available [Online]: http://www.helsinki.fi/science/optek/1995/n2/newman.txt

Ravenscroft, A. & Pilkington, R. (2000). Investigation by design: Developing dialogue models to support reasoning and conceptual change. *International Journal of Artificial Intelligence in Education, 11*, to appear.

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

Reber, A. (1995). *Dictionary of psychology (2nd ed.)*. Toronto: Penguin Books.

Riffe, D., Lacy, S., & Fico, F.(1998). *Analyzing media messages: Quantitative content analysis.* New Jersey: Lawrence Erlbaum Associates, Inc.

Rourke, L., Anderson, T., Archer, W., & Garrison, R. (in press). Assessing social presence in computer conferencing transcripts. *Canadian Journal of Distance Education.*

Sinclair, J., & Coulthard, M. (1975). Towards an analysis of discourse. London: Oxford University Press.

Weiss, R. & Morrison, G. (1998). Evaluation of a graduate seminar conducted by listserv. (ERIC Document Reproduction Service, ED 423 868).

Zhu, E. (1996). Meaning negotiation, knowledge construction, and mentoring in a distance learning course. (ERIC Document Reproduction Service, ED 397 849).