RCETJ, 2 (2), 1-14

Methodological Issues in the Content Analysis of Online Asynchronous Discussions: Unitizing, Reliability, and Latent Content

Elizabeth Murphy
Justyna Ciszewska-Carr
Maria A. Rodriguez Manzanares

Memorial University of Newfoundland, Canada

Abstract
This paper explores three methodological issues related to content analysis of online asynchronous discussions: unitizing, reliability, and manifest versus latent content. Unitizing involves balancing feasibility, reliability, identifiability, and discriminant capability of semantic versus syntactic units. Reliability is discussed in relation to differences between tasks, discussants, and number of coding decisions, as well as between coders. Manifest versus latent content contrasts observed versus intended behaviors in content analysis. For each of the three issues, the paper presents a brief theoretical overview. Each issue is subsequently contextualized and illustrated using empirical results. Finally, for each issue, the paper provides a discussion of lessons learned and implications. The paper concludes with suggestions for future studies.

Introduction

Content analysis of online asynchronous discussions has been described as “difficult, frustrating, and time-consuming” (Rourke, Anderson, Garrison, & Archer, 2001, p. 2). Likewise, Rourke and Anderson (2004) note that, while the technique is “promising,” it remains “unfamiliar” (p. 5). It is not surprising, therefore, that researchers have identified numerous methodological issues or challenges related to analyzing online asynchronous discussions.

Rourke et al. (2001) examined 19 research studies of content analysis of online asynchronous discussions. The authors identified and discussed six related methodological issues. The first issue relates to the criteria for content analysis which, as in other contexts of research, have to be met in order to ensure validity of a study. These criteria include objectivity, reliability, replicability, and systematic coherence. The second methodological issue relates to the type of research design to be adopted, and involves determining whether to take a more descriptive or a more experimental approach to the analysis. The third issue relates to the type of content to be examined, and involves determining whether, and to what extent, to focus on the manifest content or the latent content of the discussion. The fourth issue involves choosing a unit of analysis that would best suit the context and the purpose of the analysis. The fifth issue relates to the choice of software to facilitate the organization and analysis of data. The final issue involves ethical issues researchers face, such as informed consent, and the respect for and protection of discussion participants.

Fahy (2001) also focused on some common problems researchers encounter when engaging in content analysis. He suggested that these difficulties are due to “failings of both technique and … theory capable of guiding transcript analysis research” (¶ 1). One problem Fahy identified relates to the discriminant capability of frameworks used for coding. More specifically, when the framework includes numerous categories that are not mutually exclusive, assigning units of analysis to only one category cannot be effectively accomplished. Since discriminant capability is closely related to reliability, problems with the clarity of the coding categories will result in low reliability between coders. Another problem is the choice of the unit of analysis. Fahy argues that the unit has to allow for systematic and effective categorization of the data, and has to be “obvious and constant” throughout the transcript (¶ 10). For Fahy, the sentence is the type of unit that meets these criteria and is therefore the most suitable and reliable choice for content analysis of online discussion transcripts.

In a more recent study by Rourke and Anderson (2004), the authors argue that content analysis continues to be problematic due to its lack of systematicity and objectivity. They suggest an alternative approach to content analysis which involves viewing it as a form of testing and measurement. This way, methodological issues in content analysis will be considered in relation to well-established rules of test validity. First of all, this approach would involve creating a theoretically and empirically valid coding protocol. Developing such protocol requires several stages. With regards to ensuring theoretical validity, these stages are: identifying the purpose for which the coding data will be used; identifying behaviors that represent the construct; reviewing the categories and indicators to ensure adequate representation of the construct; conducting preliminary tryouts; and creating guidelines for administering, scoring, and interpreting the coding framework. Ensuring empirical validity of the coding protocol involves conducting correlational analyses, examining group differences, and manipulating the variables to see if changes occur.

The purpose of this paper is to explore in depth three methodological issues related to the content analysis of online asynchronous discussions: unitizing (Murphy & Ciszewska-Carr, 2005a), reliability (Murphy & Ciszewska-Carr, 2005b), and manifest versus latent content (Murphy & Rodriguez-Manzanares, 2005). For each of the three issues, the paper presents a brief theoretical overview of the issue. Each issue is subsequently contextualized and illustrated using empirical results. Finally, for each issue, the paper provides a discussion of lessons learned and implications. The paper concludes with suggestions for future studies.

Background

The results referenced in this paper were presented and discussed in detail in three previously published papers. One of these papers focused on the issue of reliability in content analysis (see Murphy & Ciszewska-Carr, 2005a). Another paper highlighted issues related to the choice of unit of analysis (see Murphy & Ciszewska-Carr, 2005b). The final paper considered the issue of latent versus manifest content (see Murphy & Rodriguez-Manzanares, 2005). The focus on the methodological issues drew on data which came from the content analysis of a transcript of a discussion with ten discussants. The discussants were seven graduate students and three undergraduate students enrolled in Counselling Psychology courses at a Canadian university in the Fall of 2004. The focus on the first two issues drew on data from analysis of the transcripts of all ten graduate and undergraduate students. The focus on the third issue drew only on data from analysis of the transcripts of the seven graduate students. The unmoderated discussion was part of a one-month long online module related to Problem Formulation and Resolution (PFR). The focus of the discussion revolved around how to promote parental involvement in schools. Eight tasks (prompts) guided the discussants through a process of understanding and identifying possible solutions to the problem. The first five tasks were designed to support engagement in Problem Formulation and the remaining three tasks supported engagement in Problem Resolution. An example of a prompt is as follows: *Compose and post a message in which you describe how your understanding of the problem has changed as a result of having read an article on the problem*. Each of the ten discussants was required to post one message in relation to each of the eight tasks.

Transcript analysis began with a framework for organizing and guiding the analysis. This framework relied on predetermined categories, processes, and indicators of behaviors related to Problem Formulation and Resolution (see Appendix A). The framework was a second iteration and had been developed from the literature on problem solving. It had been tested in two contexts (see: Murphy, 2004). Analysis involved comparing the text or transcript of the discussion with this framework and looking for evidence of the categories, processes, and indicators of PFR. The two main categories were Formulation and Resolution. Each category was further subdivided into associated processes. For example, one of the processes related to the category of Formulation is that of *building knowledge*, while *identifying solutions* is one process associated with Resolution. The processes themselves were further subdivided into specific indicators that operationalized and identified the types of behaviors that might be associated with each process. The framework included 19 indicators, such as *rejecting/eliminating solutions judged unworkable*.

Coding began with the choice of unit. Both semantic and syntactic units were used (see below for an entire section devoted to the issue of choice of unit). Coding involved associating each unit of analysis

first with a category, then a process, and finally with an indicator. The protocol adopted for coding limited coders to assigning only one possible code per unit. The coding was conducted by two graduate research assistants with no prior coding experience. Each coder (A and B) received prior training with the principal investigator of the study. Training involved becoming familiar with the framework and practicing coding units. In phase 1 of the coding process, the two coders first independently selected and coded the units of meaning. In phase 2, they coded the syntactic units of a paragraph. Reliability measures were calculated using Cohen's kappa (see Cohen, 1960). This paper discusses reliability results in depth in a later section (see reliability)

Following the discussion, each of the ten discussants was invited to participate in an interview with the principal investigator of the study as well as one of the coders of the discussion transcript. The purpose of the interviews was to provide discussants with an opportunity to explain why they behaved as they did in the discussion forum. Each interview lasted approximately one hour. As discussants read through their own messages, they were asked to talk about their intentions or motives when writing the message. The interviewer prompted the discussants using the PFR framework.

Issues

Unitizing

Choice of the unit of analysis is the starting point for coding the transcripts and represents a "complex and challenging" process (Rourke et al., 2001, Unit of Analysis section, ¶ 8). Units can be classified into fixed syntactic units which are determined by graphic conventions or into dynamic semantic units based on meaning. In terms of syntactic units, researchers have worked with sentences (e.g., Fahy et al., 2000), paragraphs (e.g., Hara, Bonk, & Angeli, 2000), as well as whole messages (e.g., Oriogun, 2003). In terms of semantic units, researchers have worked with themes (e.g., Henri, 1992), message maps (e.g., Levin, Kim, & Riel, 1990), or illocutionary units (e.g., Howell-Richardson & Mellar, 1996).

The choice of unit of analysis may be constrained by the identifiability, reliability, discriminant capability, and feasibility of the different units. For example, the unit must be easily recognizable or identifiable within a transcript. In contexts of computer-mediated communication, the identifiability of units can be compromised by discussants' idiosyncratic use of language and of discourse conventions. The choice of a syntactic unit such as a sentence may not be possible if discussants have ignored conventional punctuation.

In cases where it is possible to choose the sentence as the unit for coding, issues of feasibility may need to be considered. If the discussion contains a high number of sentences, yet the time available for coding is restricted, the syntactic unit of the sentence may prove unfeasible in this context. For this reason, larger units such as whole messages may place fewer demands on coding resources and thus prove more feasible. However, the larger the unit, the less likely it will be capable of supporting discrimination between the different behaviors being coded for in the discussion transcript. Challenges to discriminant capability in this case might be overcome by choosing a semantic unit. Yet, unless these units of meaning are selected prior to coding and agreed upon by both coders at that time, reliability may be compromised. The following paragraphs highlight how these issues manifested themselves, first, in coding with a semantic unit and, subsequently, with a syntactic unit.

Semantic Unit

In our study (see Murphy & Ciszewska-Carr, 2005a), we initially worked with the semantic unit of analysis. First, both coders worked independently and divided the discussion transcripts into semantic units according to their individual interpretations and judgements. To promote high discriminant capability, coders could select any portion of the text they believed to contain a complete idea. This process yielded different results for each coder. In terms of the whole transcript, Coder A identified a total of 393 units while Coder B identified 457 units. If we consider the identification of units not at the aggregate level but instead at the individual level, the range of differences is even greater. For example, in the case of Discussant I, Coder A identified 30 units in that discussant's transcript, while Coder B identified 65 semantic units. The two coders identified a different number of units for nine out of ten of the discussants. Only in one case did they reach an agreement on the number of semantic units. Both identified 48

semantic units in the transcript of the one individual. However, even though they both identified the same number of units, the actual boundaries of those units were different in most instances.
The differences in results of Coder A's and Coder B's choice of semantic unit highlights the issue of reliability. In spite of a prior training session, consistency between coders in the choice of unit was low across all ten transcripts for the ten discussants. This result illustrates one of the difficulties of working with the semantic unit. Its identification requires interpretation and judgement on the part of coders. Ensuring reliable and consistent interpretation and judgement between coders may not be possible in spite of training. If the units are different, then coding itself cannot be reliable. This will be the case even if all other factors are controlled for to ensure reliability e.g. use of a tested instrument. Furthermore, without a consistent choice of units between coders, measures of reliability such as Cohen's kappa cannot be calculated effectively.

Syntactic unit

In the second stage of our study, we worked with the syntactic unit of a paragraph. The decision to choose the paragraph over other syntactic units such as a whole message or a sentence was influenced by two reasons. One reason related to the results obtained in the first stage of the study: when the coders divided the discussion transcript into semantic units, the boundaries of these units frequently overlapped with the boundaries of the paragraph. Thus, the division of units were, in many cases, the same as the paragraph divisions. This observation indicated that the discussants tended to explore each individual idea within a separate paragraph. Another reason why we chose to work with a paragraph was the high identifiability of this type of syntactic unit in the context of this study. Discussants were not given any prior guidelines, requirements, or suggestions regarding message layout, discourse conventions, use of paragraphs, or punctuation. However, they all automatically and naturally divided their messages into paragraphs. Since the discussion took place in the context of a university course, it is possible that discussants chose to adopt graphic conventions they would normally use in a formal context of learning. This clear division of text thus resulted in easy and consistent identification of units within the discussion transcript.

The total number of paragraphs in the complete discussion transcript was 355 while the number of paragraphs within individual discussants' transcripts ranged from 29 to 55. The length of the paragraphs ranged from six to 780 words, with an average of 97.6 words. The shortest paragraphs for individual discussants ranged from six words for one individual to 31 for another while the longest ranged from 149 to 780 words. When paragraphs were longer, the discriminant capability of the paragraph as a unit may have been compromised. In fact, the discriminant capability of the paragraph was even an issue in some cases when paragraphs were much shorter. For example, one discussant wrote a 157-word long paragraph which constituted one syntactic unit. When coders were first working with the semantic unit, Coder A identified four units from within these 157 words. The identification of four units in this paragraph meant that possibly as many as four different behaviors might have been identified in this unit. This potential result highlights the discriminant capability of the semantic unit. When working with the syntactic unit of the paragraph and these same 157 words, only one potential behavior could be identified as there was only one unit. This result may suggest that, when working with the paragraph as a unit of analysis, as its length increases, its discriminant capability decreases.

We note that this result *may* suggest an inverse relation between paragraph length and discriminant capability. However, our overall results did not confirm this conclusion. The total number of semantic units identified in the transcript was 393 for Coder A and 457 for Coder B. The total number of syntactic units was 355. Final results of coding showed that the individual and group profiles of engagement in PFR behaviors were very similar with both semantic and syntactic unit of analysis in spite of this difference in the number of the semantic versus syntactic units. For example, when working with the semantic units, Coder A identified that all discussants as a group engaged in identifying solutions 31.5% of the time. When working with the syntactic unit of a paragraph, the same coder identified that the discussants engaged in identifying solutions 30.1% of the time.

In terms of the issue of feasibility, in the context of our study, the paragraph was a feasible choice with respect to the coding resources available. If we had chosen a sentence as the unit of analysis, the

number of units to code would not have been 355. Instead, it would have been in the thousands since each of these 355 paragraphs contained numerous sentences. Coding thousands of sentences would have consumed an unfeasible amount of time on the part of coders and yet may not have yielded much different results than what was achieved with the paragraph. Given the results reported above regarding the difference between the semantic and syntactic unit of the paragraph, we can conclude that results between the sentence and the paragraph may not have been much different. In that case, the more finegrained unit of meaning did not result in overall or aggregate difference in PFR behaviors. In our context, the message would have been an even more feasible choice than the paragraph in terms of the coding resources and time, as the total number of units would then have been only 80. However, since the messages in our context were relatively long, with an average of 438.2 words, they may not have effectively discriminated between PFR behaviors. Nonetheless, we cannot confirm this conclusively because we did not code any messages as a whole and also because the results reported above did not show differences based on size of the unit.

Lessons Learned

The results reported above highlight the role of context in the choice of unit. The issues of identifiability, feasibility, reliability, and discriminant capability will increase or diminish in importance depending on this context. For example, the semantic unit will be most effective at discriminating between behaviors. However, as our study illustrates, working with the semantic unit may result in low reliability between coders. In order to ensure this consistency, or reliability, it is necessary for the coders to decide on the semantic units prior to the beginning of coding. This approach requires two phases of analysis of the transcript. In phase one, the coders would need to first decide on the unit. This decision might be accomplished by a simultaneous, shared analysis for the purpose of identifying the units.
Another approach would be for the coders to independently identify the units of meaning in the transcripts and then compare results and come to an agreement on the final boundaries of the units of meaning.

Phase two would then involve coding these units. In this case, reaching an agreement on the boundaries of the semantic units would have been possible. However, it would have added an extra phase to the coding process thus requiring more time on the part of the coders. However, the need for more time then raises the issue of feasibility. Consistent identification of the semantic units would require the coders to conduct the analysis in two stages: first they would need to agree on the boundaries of each unit, and then code the units. This two-stage approach may not be feasible in all contexts of content analysis. In our context, we were dealing with almost 35,000 words with potentially hundreds of units. The time required of the coders to conduct two phases of coding would have placed demands on financial resources since these coders were being paid. This two-stage approach may not be feasible in all contexts of content analysis, particularly when the coders are geographically distributed.

The syntactic unit of a paragraph proved to be a more feasible choice in our context. It was also highly identifiable, which supported high reliability. The discriminant capability of a paragraph, however, could have been compromised, especially in the case of longer units. Interestingly, the results of working with the paragraph as the unit of analysis were similar to those of working with the semantic unit. The profiles of discussants' engagement in PFR were alike both at the aggregate level and at the level of individual discussants. We may thus conclude that, in our context, the paragraph discriminated between behaviors as effectively as the semantic unit. We did not, however, investigate the factors that may have contributed to this similarity in results. Other studies would be needed to determine whether such similarity might occur in other contexts, or whether it was a result specific only to our context.
In the case of the semantic unit, the use of a protocol that limits assignment of only one code per unit is a logical approach. When using a unit of meaning, we can assume that each unit will represent a distinct behavior because of the high discriminant capability of the instrument. However, when working with the syntactic unit of the paragraph, to enhance discriminant capability, a protocol could be adopted whereby more than one code could be assigned per syntactic unit.

Reliability

Inter-rater reliability in the context of content analysis refers to the "amount of agreement or

correspondence among two or more coders" (Neuendorf, 2002, p. 141). In cases of analysis of online discussions, the extent of agreement between coders may be reported in some studies as a simple percentage value. This value would represent the percentage of all coding decisions on which the coders agree. Rourke et al. (2001) reviewed 14 content analysis studies and found that, while only eight of them reported reliability at all, six reported it as a percentage of agreement. The appeal of this measure is that it is "simple, intuitive, and easy to calculate" (Lombard, Snyder-Duch, & Bracken, 2002).
We chose initially to calculate the agreement between the two coders as a percentage value. We first listed all codes assigned by both coders to the 355 syntactic units of a paragraph. Then, we calculated how many times the two coders assigned the same codes to units. The percentage of agreement between the coders in our study was 63%. This means that 63 percent of the time the two coders assigned the same code to the unit, and 37 percent of the time these codes were different.

The disadvantage of percentage of agreement in this context is its failure to account for the agreement that will happen by chance. Probability tells us that, in any situation, without any training or familiarity with the coding framework, coders will agree half the time even if the choices they make are random (Lombard, Snyder-Duch, & Bracken, 2002). It is for this reason that the reporting of agreement as a percentage is not an accurate approach and therefore requires alternative indices or measures. These chance-corrected measures include, for example, Cohen's kappa (Cohen, 1960), Krippendorff's alpha (Krippendorff, 1980), or Scott's pi (Scott, 1955).

The particular chance-corrected measure we adopted was Cohen's kappa (Cohen, 1960). As we reported earlier, the overall percentage of agreement between the two coders had a value of 63%. When calculated as a kappa coefficient, the agreement between the two coders had a value of .591 on a range from 0 to 1. This value represents the total agreement between the two coders across all decisions made in the process of coding for ten discussants, eight discussion tasks, and the level of the indicator in the PFR framework. According to a scale developed by Capozzoli, McSweeney, and Sinha (1999), this value of .591 indicates fair agreement beyond chance. As we discovered later, a kappa value of .7 could, in fact, correspond to a percentage value ranging from 62-90%. This range highlights, therefore, the importance of a chance-corrected measure for accurate reporting of reliability.

While the use of a measure such as Cohen's kappa to report of agreement between two coders may be more accurate than use of a percentage value, it can nonetheless by misleading in some contexts. In fact, Crocker and Schulz (2001) argue that reporting aggregate measures of agreement or reliability between coders may mask other "sources of error" in agreement. In the context of their inquiry of writing assessments, the authors argue that:

Classical interscorer [inter-rater] studies treat scorers as a single source of error, with interscorer correlations being reported as reliability coefficients …. Despite the preoccupation with interscorer consistency, other sources of error have also been recognized as important in performance assessment. The most common of these are tasks and occasions. (Crocker & Schulz, 2001, Reliability of Performance Assessments section, para. 1)

In order to identify the sources of error in our context (see Murphy & Ciszewska-Carr, 2005b), we considered the different sources of error or variables that might have affected the overall reliability value of .591. These variables include two coders, eight discussion tasks, ten discussants, 355 syntactic units, two categories, five processes, and 19 indicators of the PFR framework. For each of these variables, we calculated separate reliability measures using Cohen's kappa.
We began by focusing on calculating the agreement between coders in relation to each of the eight discussion prompts or tasks. Specifically, responses to each of the tasks were considered as a minidiscussion.

We began by considering the coding decisions made for only those messages posted by the ten discussants in response to Task one. We listed all codes assigned by both coders to the paragraphs within Task one. We then calculated the level of agreement between the coders on that particular task across all ten discussants. The kappa coefficient or measure of reliability for this Task one was .550. Subsequently, we repeated this procedure for each of the remaining seven tasks. The coefficients for

these seven tasks ranged from .349 for Task six to .664 for Task two, with a mean of .539. According to Capozzoli, McSweeney, and Sinha's (1999) scale the agreement thus ranged from poor to fair.
To investigate differences between individual discussants as a potential source of error we took a similar approach as with individual tasks. We considered each discussant's transcript as if it were a minidiscussion.

For each of these mini-discussions we calculated the level of agreement between the two coders. The highest agreement was reached for Discussant I with a kappa value of .907. The lowest agreement was reached for Discussant C with a value of .390. The mean across all discussants was .707. The levels of agreements thus ranged from poor to excellent.

In addition to tasks and discussants, we also examined the agreement between the two coders at the three different levels of the PFR framework. The framework contained two categories divided into five processes with each of these subsequently sub-divided into 19 indicators. Assigning a code to each paragraph involved making three decisions corresponding to the three levels of the framework. First, coders needed to determine whether the behavior manifested in each unit represented the category of Problem Formulation of Problem Resolution. Then, the coders needed to determine which of the five processes associated with the two main categories was represented in the unit. Finally, the coders needed to determine which of the 19 specific PFR behaviors was manifested in the unit.
Our approach to investigating these three levels of the framework as a potential source of error involved

calculating agreement between the two coders at each level separately. Agreement was calculated across all 355 units of analysis. At the most general level, when coders were only required to choose between two PFR categories, the agreement between them was excellent with a kappa value of .825. At the next level, when coders needed to determine first which category and then which process was manifested in each unit, their agreement decreased to good with a value of .724. At the most detailed level, when coders decided on the category, then the process, and finally one of the 19 the indicators, the agreement between Coder A and Coder B across the total of 355 coded paragraphs was fair and had a value of .591. Thus, as the number of coding decisions increased, the agreement between the coders decreased.

Lessons learned

The results reported above highlight the complexity of the issue of reliability in a context of content analysis of online discussions. To accurately represent the agreement between coders, a chancecorrected reliability measure needs to be adopted. Simple percentage of agreement is not an adequate index for this purpose because it does not account for the level of agreement that is expected to happen by chance. Even when a chance-corrected measure is adopted, reporting reliability for only the overall agreement between coders may not be an accurate approach. Such an approach may mask the differences that occur at the level of individual variables. In the context of content analysis of online discussions, these variables or sources of error may include the coders (e.g., their training and familiarity with the coding protocol), individual discussants (e.g., writing style), discussion tasks (e.g., clarity of instructions), or the framework used to code the data. In order to gain insight into the intricacies of agreement, all of these variables may need to be considered.

Manifest versus Latent Content

The above discussions of choice of unit and of reliability were in relation to the analysis of the manifest or observable content of the discussion. We analyzed the content of the paragraphs of each discussant's transcript for evidence of behaviors which we then matched to our framework of PFR. This process allowed us to identify and measure how and to what degree discussants engaged in PFR. However, this process of analysis of the manifest content did not allow us to gain insight into *why* discussants did or did not engage in PFR, nor did it tell us why they privileged certain behaviors over others.

To gain insight into why discussants behaved as they did required a focus, not on manifest content, but on latent content. Latent content consists of concepts that "cannot be measured directly" (Hair, Anderson,

Tatham, & Black, 1998, p. 581). In the context of content analysis, we understand latent content as the discussants' intentions and motives which are not explicitly expressed in the transcript. Analysis of latent content can provide insight into *why* discussants engaged in certain behaviors more than others, or why they did not engage in some behaviors at all. Our focus on latent content took place after we had analyzed the discussion transcripts for PFR behaviors (see Murphy & Rodriguez-Manzanares, 2005). The manifest content of the online discussion was analyzed using a PFR framework with two categories (Problem Formulation and Problem Resolution), five processes, and 19 specific indicators of behavior. Since the discussants' level of engagement in these behaviors varied, the goal of the interviews was to gain insight into their motivations for privileging certain behaviors over others. For example, the analysis of the manifest content of the discussion had revealed a low engagement in *rejecting or eliminating solutions judged unworkable*. This behavior constituted one of the eight indicators outlined in the framework for PRF, but the discussants engaged in it only 0.3% of the time throughout the discussion. Their engagement in other behaviors related to the process of *evaluating solutions* was also considerably low, with 1% for *weighing and comparing alternative solutions*, and 3% for *critiquing solutions*. One of the goals of the interviews was to gain insight into the reasons why discussants did not engage in these behaviors.

The discussants revealed that one reason why they did not engage in a critical evaluation of other people's solutions to the problem was that they did not feel comfortable criticizing the ideas of others. They were concerned that their critique might be received as harsh. Another reason for the lack of critique was the relationship among the discussants - all students participating in the discussion knew each other from class. Because the discussion did not allow for posting anonymous messages, the discussants did not feel comfortable signing messages they considered to be negative. Instead of rejecting, critiquing, or weighing and comparing solutions, they preferred to focus on the positive aspects of the solutions others proposed. It is due to these reasons that, of all behaviors associated with evaluating solutions, the one that was favored most was *agreeing with solutions proposed by others* (12% of all units).

Another goal of the interviews was to provide further insight into the discussants' engagement in *proposing solutions*. This particular behavior was the most privileged of all behaviors outlined in the PFR instrument (22% of all units). Moreover, discussants engaged in it from the beginning of the discussion, even though the discussion was designed to engage them first in understanding the problem and only after in trying to solve it. The discussants revealed that searching for possible solutions was an approach to problem-solving they naturally adopted, not only in the discussion, but in other contexts as well. This solution-focused approach was common to a number of discussants. Despite this commonality, however, different individuals adopted different strategies to problem-solving. Some discussants explained that, for them, looking for solutions meant matching specific causes of the problem with specific solutions. For others, it meant beginning with a more overall perspective. They would first look for a general solution to the problem, from which more specific solutions would follow.

With regards to the category of Problem Formulation, the discussants provided some reasons why they engaged in *redefining the problem*. They revealed that rephrasing the problem using their own voice was a strategy that helped them better understand the problem. The discussants also provided further reasons for their engagement in *agreeing with the problem as presented*. In particular, they explained why they frequently relied on their own experiences when engaging in this behavior. By using examples from their personal and professional life, they wished to illustrate their direct knowledge of the problem under discussion. The discussants' own experiences also played an important role in *identifying causes of the problem*. During the interviews, discussants emphasized that exchanging perspectives and hearing about the experiences of others was crucial to fully understand the problem and its multiple aspects. As one individual commented, "different aspects of the problem become more clear" when you begin to see how the problem relates to your own experiences.

Lessons Learned

These results highlight the role of the latent content in providing insight into the discussants' behavior in an online discussion and in helping understand why they privileged some behaviors over others. For example, in the context reported on here, discussants engaged more in Problem Formulation (53%) that

in Problem Resolution (47%). However, within the category of Problem Resolution, they privileged only two of the eight behaviors associated with the process. The discussants engaged in *proposing solutions* 22% of the time, and in *agreeing with solutions proposed by others* 12% of the time. This type of result could be used to inform the design of the PFR discussion. For example, the results suggest that discussants may need more prompting in order to be fully engaged in PFR. Individuals responsible for the design of online discussions in the future may wish to provide for equal opportunities for engagement in each of the desired behaviors. This could be accomplished by including discussion tasks that specifically require discussants to, for example, critique solutions or weigh and compare solutions. Our analysis of manifest content suggests that, in some contexts, discussants may not engage in certain behaviors unless explicitly instructed to do so.

While analysis of the latent content can provide important insights, it does not guarantee them. In our context, the analysis provided insight into the discussants' motives and intentions in some cases, but not in others. For example, the PFR framework included 11 specific indicators of behavior associated with the category of Problem Formulation. However, of these 11 behaviors, discussants privileged the following one: *accessing and reporting on sources of information*. While 14% of all units were coded for this behavior, the discussants' engagement in the remaining ten behaviors related to Problem Formulation ranged from 1% to 9%. The interviews, however, failed to provide further insight into why the discussants preferred *accessing and reporting on sources of information* over any of the other Problem Formulation behaviors.

This failure of the interviews may be due to several reasons. One reason may relate to the limitations of the interview protocol. In order to identify their own motives and intentions, discussants needed to engage in a form of metacognitive reflection. Since interviews did not reveal the discussants' motives for all of the PFR behaviors, we may speculate that the interview questions may not have adequately supported engagement in metacognition. Another reason may relate to the retroactivity of the interviews, which were conducted a month after the discussion. It may have been challenging for the discussants to recall what they were thinking when writing their messages.

Conclusion

This paper presented a synthesis of three studies that illustrated and explored methodological issues related to the content analysis of online asynchronous discussions: unitizing (syntactic versus semantic unit), reliability, and manifest versus latent content. Each of these issues was discussed in relation to empirical results from an analysis of an actual online discussion. The issues considered in this paper highlight the complexity of content analysis of online discussion transcripts. They confirm other researchers' perspectives that content analysis is indeed "difficult" and "unfamiliar." Our discussion, it is hoped, has made this area somewhat less difficult and more familiar. In this regard, we noted a number of lessons that might assist future content analysts.

In terms of unitizing, we found that issues of identifiability, feasibility, reliability, and discriminant capability will increase or diminish in importance depending on the context of content analysis. However, we only worked with the semantic unit and the syntactic unit of a paragraph. Other researchers may wish to adopt a different syntactic unit of analysis, such as the sentence, and see how those four issues manifest themselves in that context. In terms of reliability, we found that simply calculating an overall reliability measure for all variables does not adequately represent the range of agreement between coders.

Therefore, a more fine-grained measure of reliability might be necessary. While we identified a range in reliability measures between different variables such as, for example, different participants, we did not go a step further in terms of identifying why certain variables produced higher or lower reliability measures than others. Other studies might consider the factors that contribute to the variety in agreement at the level of individual variables. In terms of manifest versus latent content, the interviews revealed some insights about certain behaviors of the discussants. They did not, however, reveal insights about engagement in all of those behaviors. For this reason, researchers may wish to adopt other interview protocols. They might, for example, engage discussants in reflection about their intentions at the time they are composing their messages. One approach might involve a verbal expression of their motives in

the form of a think-aloud procedure (Willis, 2005). Another approach might involve a written expression in, for example, the actual body of the message.

This paper was limited to consideration of only three issues. In addition, these issues were considered within the context of analysis of one discussion only. Other contexts of analyses might shed more insight into these issues. The discussion that was analyzed in this context related to Problem Formulation and Resolution. Other contexts might highlight discussions with a different focus in order to see if the issues manifest themselves differently depending on the context. We worked with a relatively small corpus of data given that there were only ten participants in the discussion. Larger groups might provide a different perspective on the issues. Moreover, we took a deductive approach to analyzing the content of the discussion, whereby we matched the discussants' behaviors with pre-existing categories. More inductive approaches, such as a grounded theory approach, might result in a different manifestation of the issues of unitizing, reliability, and manifest versus latent content, and might highlight a range of issues not identified in this paper.

Acknowledgements

The study reported on in this paper was funded by the Social Sciences Research Council of Canada. This paper was presented at the 6th International Educational Technology Conference, 2006, in Famagusta, North Cyprus. It was published in the conference proceedings:
Murphy, E., Ciszewska-Carr, J., & Rodriguez, M. A. (2006, April). Issues in the content analysis of online asynchronous discussions. *Proceedings of the 6th International Educational Technology Conference, Famagusta, North Cyprus*, 1231-1240.

References

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*(1), 37-46.
Crocker, R., & Schulz, H. (2001). *Design of a generalizability study for SAIP assessments.* Report submitted to the Council of Ministers of Education, Canada.
Fahy, P. J. (2001). Addressing some common problems in transcript analysis. *IRRODL Research Notes, 1*(2). Retrieved January 26, 2006, from http://www.irrodl.org/content/v1.2/research.html#Fahy
Fahy, P. J., Crawford, G., Ally, M., Cookson, P., Keller, V., & Prosser, F. (2000). The development and testing of a tool for analysis of computer mediated conferencing transcripts. *Alberta Journal of Education Research, 46*(1), 85-88.
Hair, J. F., Anderson, R. E., Tatham, R. L., Ronald, L., & Black, W. C. (1998). *Multivariate data analysis* (5th Ed.). Upper Saddle River, NJ: Prentice Hall.
Hara, N., Bonk, C. J., & Angeli, C. (2000). Content analyses of on-line discussion in an applied educational psychology course. *Instructional Science, 28*(2), 115-152. Retrieved January 26, 2006, from http://crlt.indiana.edu/publications/journals/techreport.pdf
Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *Collaborative learning through computer conferencing* (pp. 117-136). Berlin: Springer Verlag.
Howell-Richardson, C., & Mellar, H. (1996). A methodology for the analysis of patterns of interactions of participation within computer mediated communication courses. *Instructional Science, 24*, 47-69.
Krippendorf, K. (1980*). Quantitative content analysis: An introduction to its method*. Beverly Hills: Sage Publications.
Levin, J. A., Kim, H., & Riel, M. M. (1990). Analyzing instructional interactions on electronic message networks. In L. Harasim (Ed.), *Online education: Perspectives on a new environment* (pp. 185-214). New York: Praeger Publishers.
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.
Murphy, E. (2004). Promoting construct validity in instruments for the analysis of transcripts of online asynchronous discussions. *Educational Media International, 41*(4), 346-354.
Murphy, E., & Ciszewska-Carr, J. (2005a). Contrasting syntactic and semantic units in the analysis of online discussions. *Australasian Journal of Educational Technology, 21*(4), 546-566.

Murphy, E., & Ciszewska-Carr, J. (2005b). Identifying sources of difference in reliability in content analysis of online asynchronous discussions. *International Review of Research in Open and Distance Learning, 6*(2). Retrieved January 26, 2006, from http://www.irrodl.org/content/v6.2/murphy.html
Murphy, E., & Rodriguez-Manzanares, M. (2005). Reading between the lines: Understanding the role of latent content in the analysis of transcripts of online asynchronous discussions. *International Journal of Instructional Technology and Distance Learning, 2*(6). Retrieved January 26, 2006, from http://www.itdl.org/Journal/Jul_05/article03.htm
Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage Publications.
Oriogun, P. K. (2003). Towards understanding online learning levels of engagement using the SQUAD approach to CMC discourse. *Australian Journal of Educational Technology, 19*(3), 371-387.
Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education, 12.* Retrieved January 26, 2006, from http://communitiesofinquiry.com/documents/2Rourke_et_al_Content_Analysis.pdf
Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development, 52*(1), 5-18.
Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 17,* 321-325.
Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Appendix A

Second iteration of an instrument for identifying and measuring PFR in an OAD (Murphy, 2004)
Process Indicator Example
Agreeing with problem as presented in OAD: *…there is a problem with getting students to speak French in the classroom, or for that matter to get the teacher to speak French in the classroom.*

Specifying ways in which problem may manifest itself: *The reality is that students will use English, or their first language to communicate as often as possible.*

Redefining problem within problem space: *Perhaps the ultimate question really should be: when should L1 be used in the classroom, what contexts make it acceptable and beneficial to speak English instead of French?*
Minimizing and/or denying problem: *I would argue that you can indeed use English in theCore French classroom.*
Identifying extent of problem: *It seems to me that this issue of French/English use in the classroom will be one of the biggest challenges we will face as teachers.*

Identifying causes of problem: *My understanding of the problem is that core French teachers are unsure of how much French to use because they don't know how much their students will understand.*

Defining problem space

Articulating a problem outside problem space: *I believe it is true that non-English speaking children are losing their mother tongue through the education system. Look at the focus of our ESL programs.*

Identifying unknowns in knowledge: *How can we reach those students who have below grade level skills, and provide them with some understanding of the target language?*

Accessing and reporting on sources of information: *According to the author, pupils should be allowed to use English between themselves while working in teams.*

Identifying value of information: *This article was not effective in teaching me about this problem.*

PROBLEM FORMULATION

Building Knowledge

Reflecting on one's thinking: *Once again, the negative view I previously had on this problem is becoming increasingly more positive.*

Proposing solutions: *I feel teachers need to use French more if they expect their students to use it.*

Identifying Solutions

Hypothesizing about solutions: *I believe that if a teacher were to make mistakes and correct them in front of a class, it would ease the students' minds about making mistakes themselves and enable them to correct themselves as well.*

Agreeing with solutions proposed by others: *I agree strongly with participant 6's views. Especially for Immersion students.*

PROBLEM RESOLUTION

Evaluating Solutions

Weighing and comparing alternative solutions: *I sincerely believe that using the target language 100% of the time creates a stagnant environment for learning. On the other hand, too much use of English would only serve to 'baby' students.*

Critiquing solutions: *While I agree somewhat with participant 3, I think some students at lower levels may become too frustrated when trying to learn the language when a teacher uses only French.*

Rejecting/Eliminating solutions judged unworkable: *I don't think it is right to start the year off with a solid plan of attack.*

Planning to act: *Personally, I have decided to speak English the first day of classes.*

Acting on Solutions:

Reaching conclusions, or arriving at an understanding of problem: *The methods which all of these sources have suggested prove that language use in the classroom is a major problem, but is also easily mended with use of the proper tools, and creativity*