Matters of Fact, Opinion, and Credibility: Distinguishing Stochastic from Substantive Information in Texts
Carl Roberts

One thing my colleague Roel introduced was the distinction between representational and instrumental approaches to analyzing text. This is an idea that Gil Shapiro developed in a chapter on text analysis in a book that I edited a number of years ago. He is so humble that he never takes credit for it. Yet I think it is a really neat idea. His idea has to do with who you consider to be the expert. On the one hand, there is the approach of the researcher who considers himself an expert – the one whose perspective is being applied, who developed the software for getting at that perspective, etc. The other approach – the one that Roel and I have been making use of over the years – has to do with what Gil called a representational approach to analyzing texts. The objective here is to represent the perspective of the person being studied, to get at the semantic information that Klaus was talking about earlier. And so it is with this latter perspective that I approach my talk at this conference. I see myself as someone interested in figuring out what is going on in the minds of these people at ANES. So instead of speaking about me and what I do, my purpose is to outline what they might consider to be a reasonable approach for dealing with ANES concerns regarding concrete, responsible ways for dealing with verbatim responses to open-ended survey questions. I shall end my talk with a recommendation.

Let me start with an overview. I am going to start off, by the way, with an assertion that even among text analysts is a controversial statement. An assertion that one is encoding "the" content in text is, for me, a nonstarter. Let me be very clear on why I think this is the case. I am certainly not going to name names. However, there are people who really believe that there is something called "the content" in text. It is a just a matter of finding the right measures for it. Next, I shall be getting at methods of encoding text. What I would like to do is to think about text analysis methods in terms of the questions that are being asked of the texts. First, let's consider what I started talking about here – the idea of expertise. Let's think about whether the investigator's expertise is being assumed, or whether we are assuming that the respondent is the one who is the expert and that I am a humble person who is trying to "get into the head" of this other. This latter would be the representational approach. With the representational approach, the expert would be the text's author; with the instrumental approach the expertise resides with the investigator. Next, our inferences may be about theme prevalence–what was referred to in Roel's presentation as frequencies in a data matrix. (With such data one can do co-occurrence analyses and so on.) You can also look at relations among themes, and then do a semantic analysis or even a network analysis. I am not going to be talking about a network analysis now, although there might be an opportunity to do that at a later time. Finally, I shall close with a fanciful recommendation.

Assumption 1: Again, this is not a statement unanimously agreed upon by text analysts. I presume that *language is NOT a neutral medium through which "content" is transmitted.* Language is something more than that. Linguistic expressions are simultaneously manifestations of perspectives and propositions regarding sensory experiences. I am going to take a little poll here. (This is kind of like "the William Rehnquist question on the ANES surveys" to find out how much political knowledge you have.) "How many people here have heard of Jacque Derrida?" . . . I'm so excited. Jacque Derrida is a French postmodernist. Derrida came up with the concept of "the hinge" in his book, *Of Grammatology.* Although in the book he attributes the concept to someone else, his whole philosophy "hinges" on this concept. The idea here is that there are phrases – things people say. When someone utters one of these phrases, they are linking the phrase into two things: expectations regarding a set of rules (namely, a grammar, a language, a perspective) and expectations regarding the phrase's propositional content (i.e., regarding a sensory experience that one might have). The idea is that whereas some articulations may be more or less grammatical than others, some sensory experiences will have greater *différance* with the concepts being conveyed.

Thus when we talk about linguistic expressions, we refer to things with dual lives – things that sort of float between grammar and sensory experience. Then the issue becomes, "How am I going to

understand a particular expression?" Well, okay, we might start by asking ourselves, "Whose grammar (or perspective) and whose sensory experiences are we talking about?" Let's first assume that the investigator is the expert regarding her own sensory experiences. Focusing only on the lower half of Derrida's hinge, this involves considering the investigator's sensory experience and her expectations regarding the sensation of "Chief Justice." Now, does "this empirical sensation" count as a correct response to the question, "Who is this fellow, Rehnquist?" It may be the case that the coder has such a restrictive definition of what constitutes a correct answer, that the coder's sensory experience of "Chief Justice" just doesn't fit her expectations, and she decides, "No, it does not count."

On the other hand, we might focus on the upper half of Derrida's hinge and consider the expectations that our investigator has regarding "what is a grammatically correct reference to a senior job in a government position." Here we are referring to the perspective that the researcher brings to bear on the phrase, "Chief Justice." Maybe the phrase would be counted as "sufficiently grammatical" to count as an instance of a correct articulation of what Rehnquist's job is.

Next, let's consider the respondent to be our expert. I'm going to give a bit of a flippant illustration here. (Yet actually other relevant illustrations have been mentioned during our previous discussions as well.) If the respondent, as expert, is asked, "What is Rehnquist's job," and she replied, "Well, the other day I saw him sending an appeal back to a lower court." Your reaction would surely be, "Well, sure that is the nature of his job." More seriously, however, there are often times when we do think of the respondent as the expert. This usually occurs when the respondent has life experience that we just cannot tap into directly. For example, we might ask them about crimes in their neighborhood. Or we could ask them about who criticizes whom in the family unit. Although we have no direct access to these sensory data, we can ask respondents about their experiences and use them as our experts as we seek to understand the relations inside of their families.

Okay, then what about respondents' expertise regarding their personal perspectives? Remember the comment during a prior discussion about Cheney being the most dangerous person in America? Well, what about a response to the "what is Rehnquist's job" question like, "a job done worse than by any of his predecessors?" Would this not be a legitimate answer to the question, namely that his job is one that has been done pretty terribly?

Notice here that we have reviewed four legitimate answers to the Rehnquist question – answers that vary according to whether or not we ascribe expertise to the perspective of the investigator or the perspective of the respondent. It seems to me that this is an a priori question. Sometimes deciding the locus of expertise is pretty straightforward: If we are trying to evaluate someone's political knowledge, we might ask them about a number of people in the government – what positions they hold and so on. In this case, we are the experts. Our reaction to the respondent is not, "Wow! I didn't realize that Rehnquist was the Chief Justice." Of course not. I am the expert, and I am testing whether or not the respondent knows if Rehnquist is the Chief Justice. The key here is that it is going to be a surprise when I find out what an expert tells me, whereas when I am the judge this is not going to be the case.

What I am suggesting is that we think more about the open-ended questions we are asking. It is not that we are revealing content through some sort of black box (at least at in the encoding stage of the research process). We might like to think that there is this universal black box for revealing THE content that is out there. A more fruitful revelation might be to ask ourselves, "What are the questions that we are bringing to the text?" My argument is that there are two key questions:

Question 1: As investigators are we the experts, interested in identifying the words respondents' are using or in diagnosing how they use them, or are we novices interested getting factual or attitudinal information from respondents? That is, do we presume the respondent or ourselves to have the appropriate perspective for interpreting the text at hand? Well, one might ask, why don't I just talk about what it means commonsensically – as if there is some general audience out there? This is an argument that I have heard – one that I had to think about for a long time to come up with a really

good example. Here is what I came up with: Here is a statement by a politician at one point in history. (I will give you the background in a moment.) The statement was, "When our Prime Minster negotiates, his first offer as his last offer!" If I am going to encode this information, I must first understand what it means. It sounds as if the Prime Minister is a hard bargainer, until you find out that it was said in 2001 by Ariel Sharon regarding Ehud Barak who was Israeli Prime Minister at that time. Given this context the reference to Barak gains a whole new meaning for us. The idea is that we have more than the statement that is being said. The meaning of the statement (i.e., Sharon's intended meaning) was if Barak makes an offer he gives away the store. It is not that he is a hard bargainer. So what is the general audience here? If I am coding according to the audience's perspective then answering this question might be problematic.

Another illustration is from a similar time in history. Here is a bumper sticker that came out around 2001: "Don't blame me, I didn't vote for him..., I think." It was a Florida bumper sticker. As non-Floridians, we might interpret this as suggesting that Floridians are stupid. But then I got to thinking: Since this is a bumper sticker on a car from Florida, how does the car owner from Florida interpret it? Very likely she doesn't think that Floridians are stupid. Instead her meaning is probably, "We have a bunch of incompetent officials in our state." Note that we again have different audiences here. It is the same statement but we are interpreting it differently. So the bottom line here is that audiences have multiple perspectives. If we want to have a singular target for our encoding, we have to apply something that is (at least theoretically) singular. I submit that this must be either the perspective of the person whom we are interviewing or our perspective that we are imposing in some way on our interviewee's words.

If we have an investigator as our expert – and this is the beginning of my thinking about answers to problems that arise when encoding open-ended survey questions – if I am an expert (and thus if I have a perspective on, for example, what I understand it means for someone to be knowledgeable of American politics), then I ought to be able to come up with an algorithm for encoding respondents' answers to my questions and do an awfully good job at it. After all I am the expert; I am the one imposing the rules here. Accordingly, we can use techniques such as fixed dictionaries, disambiguation routines, machine learning, and so on. We can talk about the General Inquirer – this automated encoding software has been around the longest. The program was tailored to classify text into Osgood's semantic dimensions. And then there are machine learning algorithms. We may begin with different definitions of what constitutes similarity. We may start classifying things into different categories. But if we ourselves are the ones who are the gold standard for understanding what "true classifications" are, we should be able (if we are self-conscious enough about our perspective) to come up with a consistent way of operationalizing our encoding procedures in some sort of software, it seems to me.

We also have more sophisticated types of analysis when our investigators are experts. Special purpose software exists with parsing algorithms, not only for classifying words into semantic categories, but also for identifying the relations among the words in these categories. For example, Lou Gottschalk (on the west coast) developed software for handling encoding of psychological states. He is the psychologist; he is the expert. His understanding about what constitutes verbiage from somebody who is anxious or hostile – something about which he is an expert –is something he has operationalized into his software which, by the way, does precisely this sort of parsing. Then there is Phil Schrodt. In a number of the conversations between presentations, people have talked about his program, TABARI (its latest version). TABARI was developed to parse Reuter's new clips into data on who did what to whom (and so on) among different world events.

On the other hand, when our respondent is the expert, one might ask a question like, "What do you think was the most important issue facing the United States in the last four years?" For me this is a nice illustration of a question in which we have a respondent who is the expert. In such cases, we researchers have to be really humble. The problem isn't going to be so easy as saying, "Well I am the expert, so I can come up with a way to operationalize my perspective." Instead, I have to try to get the perspective of this person – the respondent. Then what we get into is what I think Klaus so elegantly talked about, namely the issue of semantic validity. Remember his diagram with all those circles over

here? And here we have THE (singular) world of the researcher. Now this latter world does seem to be relevant when we are talking about the researcher as the expert. But when we are talking about our respondent (out here), it could be any of a number of these worlds that is relevant. And it is up to us when encoding their responses to try to get into their worlds. What is a "top honcho?" This is a legitimate response, if you think about it – "the top honcho of the Supreme Court." Sounds like maybe the respondent knows what she is talking about. For such cases we might consider software assistance – the sorts of computer aids Roel was referring to. Key-Word-In-Context aids are particularly helpful in this process. With them you can take a look at the various terms and phrases that people use, and place them in context (i.e., with preceding and trailing words) that you can see on the computer and that can help you make encoding decisions. Of course, we can also leave a "paper trail." This can be a virtual paper trail enabled by your software program where you can specify the rules you are using when basing coding decisions on idiomatic phrases, contexts, and the like. In developing a dictionary online you should make sure that you handle the same phrases the same way throughout. That, of course, is what contemporary software aides can assist you in doing.

Kristin Behfar will be speaking tomorrow. If I understand her "concept mapping method" correctly, she provides an alternative to having the categories of the expert investigator imposed on the individual. Instead, representatives of the respondent population are asked to develop a set of semantic categories that the researcher adopts for later analysis. Here we have another approach to analyzing texts. But the key here that we are now considering the respondent to be the expert. These are the a priori questions (regarding who is expert) that we need to think about.

My software is called TCA, Textual Content Analysis. Roberto Franzosi's new software is called QNA. These are all products that allow you to get into the semantic relations that exist among the words in one's texts. Up here (first two bullets) we may be developing a set of mini-categories, here (3rd bullet) we would be developing ways in which encode the relations among the words themselves. Now to my second key question:

Question 2: "Are the inferences to be made about theme prevalence or theme relations?" Are we interested in the conditions under which some themes are more prevalent than others, or are we looking for those relations that I was just talking about – relations that might be encoded using the last kinds of software? The software and methodologies mentioned in the previous two slides refer to this distinction. So I'm just sort of reiterating here. What I mentioned at the bottom of the last two slides has to do with relational approaches to encoding texts. If we are looking at prevalence among themes – thank you Nigel for making this observation in the preface to your talk – what we are dealing with is data reduction. I guess that a lot of people don't recognize that data reduction is a lot of what we are doing in content analysis. It certainly is central.

In just coming up with our thematic categories, we take our first big step into data reduction. And then we may legitimately draw inferences about the prevalence of these themes in various types of texts. What can we do with a bunch of word counts? We can talk about prevalence. Yet co-occurrence analysis is something that people working with word counts oftentimes do. I want to talk now about a methodological issue that is a concern to me – not so much that people do co-occurrence analyses as what they do with them. In a co-occurrence analysis one obtains correlations among word counts, often ending up with scatter plots in which Euclidian distances among various clusters of these concepts are displayed. These plots are kind of interesting. They show which themes tend to be used together in the same text blocks and which ones aren't. You can do a lot of interesting work on agenda setting with such analyses by examining what kinds of clusters precede others and which co-occur at what times. The problem comes when you start referring to "word-frequencies that are correlated" in terms of the words' semantic relations to each other. Co-occurrences may be found within text windows: within large blocks of text or within exceedingly small windows. In all cases, the problem remains that you are committing the ecological fallacy when semantic relations are inferred from co-occurring meaning categories. Remember the ecological fallacy committed by Durkheim when he argued that provinces with high suicide rates and large proportions of Protestants afforded evidence that the Protestants were committing suicide (rather than, for example, that the Protestants were driving the Catholic minority to self-destruction)? The key here is that if we wish to draw

inferences about clause-level relations – that is, about relations among specific themes – based upon correlations among their frequencies in a larger text blocks (without directly going in and encoding them ourselves), we've got problems. These are problems that one has to be able to defend in publishing the work that you do. The bottom line is that if a research question deals with grammatical relations among words, these relations really need to be encoded at the outset.

How themes are related. National cross-sectional surveys might be used to reconstruct facts that respondents may have observed. This is a point that I mentioned earlier. For example, one might investigate who criticizes whom within the family unit. Here notice that what we are doing is "getting at the facts." We are trying to reconstruct facts. A lot of Roberto Franzosi's research has to do with what it is that happened during strike activities in Italy at certain historical periods. We are talking about historical research here – understanding who did what to whom, what was the timing of events, what were the associations that people had, and so on. Examining historical documents we can encode relations of who did what to whom at various times and see how the actors and the actions changed over time. Extrapolating to information that we might get in a national survey, we could get information on who criticizes whom, for example.

What happens more commonly in national cross-sectional surveys is that researchers seek respondents' perspectives as reflected in such topics as "the most important issue facing the United States." My own work in semantic grammars investigates cultural perspectives. (If I have time, I will talk about that in a minute.) I am first going to give you an idea of what we mean by semantically encoded data, and then talk a little bit of the trade-off we have when we do research on "the facts" (that is, on who did what to whom) versus when we seek information on perspectives that might be gleaned from relational data.

In the 1960s a follow-up question was asked to the "most important issue" question on the ANES survey. The follow-up question was, "What would you like to see done about this problem" – this most important problem. We might have, for example, a response to this like, "We need to keep jobs in the United States." Note that it is going to be really hard to encode this without some kind of relational information. For example, we might think of this as an expression of "we" (the subject), "keep" (the verb), "US jobs" (the object), and then we have the modal auxiliary verb, "need" (or "is necessary"). So I have made a matrix – the code for "we" might be 11, "keep" would be 35, "US jobs" 64, "necessary" is 1. Note that here we are avoiding the ecological fallacy, because we have identified these relationships in the text and we have encoded them within our data matrix.

Here is the next point I want to make: If you have semantically encoded data, you aren't gong to be able to both analyze the facts and get at the respondent's perspective. So here is the argument: In conveying facts (i.e., the facts respondents have witnessed), differences among the perspectives respondents use in interpreting the facts are noise. If I have someone who believes, "Husbands should tell wives what to do but not the reverse," and the facts are that the husband told the wife what to do, no criticism will be reported. But if she said something back to him, then that "fact" might very well be reported as her having been critical. Even though you are asking who criticizes whom within the family unit, the ideological perspective that is being used is detracting from objective observations about where criticisms are coming and going. On the other hand, if we are looking for expressions of perspectives amidst relational information, differences in the facts of people's lives (for example, whether they just lost their job or got a divorce) are going to alter their renderings of their perspectives. So what we have here is a trade-off: *respondents' perspectives introduce noise into reports of the facts, and facts introduce noise into renderings of perspectives*. This is an a priori problem – a problem that calls for an a priori decision on whether one is to study facts or perspectives. One cannot study both simultaneously.

Here is a quick summary: Two key questions basic to any text analysis are, "Who is going to be the expert?" and "What are we going to make inferences about?" Answers to these questions impose substantive structure on one's data – structure independent of stochastic variation among responses. Since these questions have no correct answers, there is no single best method of encoding text. The bottom line here is that when presuming investigator expertise, respondents' meanings for their words

are ignored. Although this approach lends itself to automation, it is restricted to inferences relevant to the investigator's perspective. When presuming respondent expertise, texts may afford data either on these respondents' perspectives or on the facts they have experienced. One of these always introduces noise into findings about the other.

Finally, inferences about semantic relations are only legitimate if the relations have been encoded into one's data. These are the points that I wanted to make.

(Do I have time to show a couple of additional slides here?) This is going to be my recommendation: Since we have a relatively small community of scholars who do text analysis, why not give them access to the verbatim transcripts? (Although I think you have heard this a number of times, I am going to reiterate it.) These would be provided only to the subset of the scholars who have interest in the ANES survey data which, of course, is going to be a subset of all text analysts. Respondents' anonymity could be ensured under penalty of losing one's firstborn child. Also, we have been talking about removing potentially identifying terms like "Maserati," "General Motors," and other proper nouns from the data. Under these conditions you aren't going to have any problems with privacy, it seems to me. You can make promises from researchers as stringent as you like. I was also going to talk about modality analysis, but only if people want me to spend some of the next 15 minutes on it.