CATA (Computer Aided Text Analysis) Options for the Coding of Open-Ended Survey Data
Paul Skalski

Thank you Jon and Skip and everybody else. I wrote some textbook material on computer aided text analysis or assisted text analysis for the Content Analysis Guidebook back when I was a graduate student at the Cleveland State University. I have kind of kept up with it a little bit. I maintain a website that lists all the programs. This conference has gotten me back into it. I want to get more into it after this is over. Thanks a lot to everyone who has presented so far and involved in this.

A little background on computer assisted text analysis, in 1960's the first programs that people know about is the General Inquirer created by Phillip Stone and his colleagues at Harvard. This is the program that really pioneered the use of computers for content analysis of textual information. This was a big undertaking, supported by a lot of grant funds. National Science Foundation had some interest in this for surveillance and security perhaps. I was thinking about this on the way here – the government probably has some great computer text analysis options. I don't think we will get those any time soon. I will talk about what is out there right now. This one, the General Inquirer, was actually a mainframe computer program. They had to go to Harvard and use theirs. To show you how far this technique has come in terms of accessibility – I was going to show you that the original General Inquirer appears online now as a web Java program that you can use instantly. You can do an instant computer text analysis with 182 coding categories. The dictionaries, I will talk about those in a moment. I just looked at the site and it is down. I think I can go to a version of it. It just came back. This is how easy it is. You just paste your text in here. It will give you instantaneous results. This shows how easy it has become to use computers that deal with a lot of textual information. I just have to paste in the instructions here. It also shows how stupid computer text is – it is going to analyze this on a number of features and it doesn't recognize it as the instructions for the program. What this gives you then – this will code, this is the dictionary, these are concepts that are part of this general Inquirer system. There are things here – language that indicates weakness, power, virtual, positivity, negativity, etc. it will give you a frequency and what percentage of the text. That is just one example of computer text analysis, an easy one that you can do on your own.

My background again, I worked on content analysis. Dr. {?} also has a pretty good book out on content analysis. I compiled a list of all the programs – 8-10 years ago there were 20 programs identified most were for Windows operating system, most were for commercial programs where you had to pay some amount. I will get into the costs at the end. Now I have been maintaining a list of these programs on the Content Analysis Guidebook online which is the website companion to this book. As you can see there are some newer programs that I put separate just to highlight them. There are qualitative analysis programs if you are into that. Other programs are hard to figure out. Also some video/audio analysis programs are available.

The point here today is that text analysis programs – there are lots of programs as you can see. A friend of mine just counted the other day – there are 30 programs, 24 are currently active today which means that some of the links go down from time to time if the programs are no longer supported or no longer available. There are a number of programs available right now. There was something that I heard at this conference that I want to add when I get back. There are quite a few different programs. You might think that all of them do different things. Really there is a lot of overlap to what these programs do. As I show in the book it basically boils down to several main functions. I am going to talk about some recommendations for analyzing open-ended survey data in a little bit. I just wanted to highlight some of the features of these programs. In my presentation I want to make this clear – I am not promoting this technique as a valid scientific technique. I'm not trying to promote any particular program; I just want to show what can be done and what some of the options are. We talked about the theory yesterday. I just want to talk about what can be done today.

The features of these programs can do several functions. Frequency and alphabetic output. You can get a list of all the words that appear in the text or a group of text. You can get those words alphabetically. Multi-unit data file output; I will explain that in a minute. KWIC or concordance – that is

where you have key word and context. If you have some words that you are interested in you can see how other words fit around those. Dictionary coding – that is really a bread and butter of CATA. This is where you apply some kind of concepts and interests to a text and code for how much those things occurred. Special analyses – I will get into those.

In terms of open-ended survey data – what would be really useful for those – I highlighted the ones I thought would be best. Frequency output I think is valuable; looking at how many times certain words appear. Also special analyses I will get into those, there are a couple that I think are kind of interesting. The big ones that are important for open-ended survey data analysis are multi-unite data file output and dictionary coding. I will explain those in more depth.

We have a little bit from Roel and Carl yesterday but I wanted to say a few more things about it today. Multi-unit data file output basically displaying output in case-by-variable form. You have each case correspond to a person and you are giving an answer to an open-ended question, you have variables that you code for. It could be positivity or negativity, some of the things like in the Harvard dictionary. This is important if you want to match up computer coded open-ended data with closed-ended data or just to convert it over to numeric form. Again, I will talk about this later. You loose some of the richness of course when you convert open-ended data to numeric data but for statistical analyses that you might want to do, it is obviously important. Most of the programs support this, multi-unit data file output, but some don't so that is something to look for.

Dictionary coding is the big thing I wanted to say a few things about right now. We talked about this already; I'll say a few things about it though. The big thing issue here is the issue of measurement. When you dictionary coding of text it is an issue of getting from conceptualization to operationalization. You have concepts of interest. You might be interested in things like positivity and negativity – how positive is the text. Operationalization is when you go beneath these concepts of positivity and indicate things that would indicate positivity, list things that would indicate positivity; words that would show that. or it could be something as simple – if you just want to analyze some of the open-ended data from ANES – what issues are most important today – you might have concepts similar to what David talked about too – issues like economy or homelessness or jobs or whatever. Also if you want to code about a particular candidate the concepts would be things like experience, performance, ability, record, personal qualities – those are the concepts in what you want to do with dictionaries is to operationalize it by coming up with terms, key words that would indicate those concepts. That is the measurement issue with dictionary coding. The two primary options here are standard dictionaries are dictionaries that are already in a program or already out there for people to use. And also custom dictionaries, this is probably the more likely one that you will use.
What you are doing with standard dictionaries is that you trust someone else's operationalization for your conceptualization, if it is even there. But in a lot of cases you might be interested in something that no one else has done. You have to create your own custom dictionary. I will go over some of the procedures for that.

A third option that I will mention briefly is emergent coding which doesn't require a dictionary. This is where a program will just go through and extract information about text, how words co-occur and things like that. I will talk about these in more detail now.

The first one is standard dictionaries – these are measures built into a program, or, trusting someone else's operationalization of your concepts of interest you might have or something that you might not have thought of but you think it would be interesting to see how people score on these concepts. Standard dictionaries, some of them seem pretty impressive if you can measure positivity or negativity or the Harvard dictionary. I think one of the problems with these that I have noticed – most CATA programs do not reveal the full dictionary information. They say this dictionary will measure optimism; it is going to measure pessimism and all kinds of concepts. But they don't reveal the full information because they don't reveal all the terms that go into it. Even when they do they don't reveal the full algorithm – how they are weighted, how things like negation are dealt with and things like that. There are some questions about this technique. It raises validity questions. It is almost as if you are reviewer for a journal and someone sends you this study and when they get to the measurement section they

say they are measuring these things but I'm not going to tell you how I did it, just trust me. That is what a lot of these programs do. I am pretty skeptical about that. In some cases if you can see all the words that go into it – the more information you can see – the devil is in the details – if you can see the details in these dictionaries then you trust them more. There are validity problems. Here is an example of that from a program that is actually my favorite of the ones I will be talking about, at this point I would recommend for coding open-ended survey data, a program called WordStat which is added onto a statistical program called SimStat. This program includes some sample dictionaries – the thing about them is that they are for demonstration purposes. They weren't intended for actual studies. The demonstration dictionary is for coding personal ads; looking for things in personal ads that people might have. One of the dictionaries that is in this program is called Sports. Here are the terms. The concept of sports is operationalized as these things. You notice any things here? It is pretty obviously missing some key sports, football, basketball, hockey jumped to my attention. It almost seemed like something got cut off in the middle. If you look closely at these dictionaries you might find things like that. You see some of these and really wonder. That is one of the problems of standard dictionaries. But again, it could be a basis for a custom dictionary so there is some use and some of them are pretty good.

Custom dictionaries, for this you have to come up with your own operationalization. This is a pretty difficult process for certain types of concepts or types of questions you might have. I have a little advice on that. The thing about it – more control and confidence because you know what is going into these custom dictionaries. The problem here, we talked about this a little bit yesterday, disambiguation – you can have a word like fine. Fine could mean a traffic violation or fine wines or fine linens it could be used like feeling fine. Words like fine, how do you disambiguate that through a computer program? You have to come up with some kind of system for that. I don't think any of them do that very well. Negation is another one. If you are measuring patriotism and one of your operationalization said any word related to patriotism and someone says that they are not patriotic – just on a simple level if the program goes through and counts those words it would count that as patriotic. Negation is really an issue that has to be dealt with. Colloquial speech, we talked about that, figures of speech, computers can't really recognize that very well either. There are some issues with these dictionaries. These could also apply to standard dictionaries of course. Being able to see helps with this. I don't think it is a fatal problem with these, I have never necessarily gone through to see how often negation appears in open-ended survey results. I think it is a problem but not a fatal one overall.

This is how we get around some of these issues with dictionary creation. Of course the big thing is identifying all the words consistent with conceptual definitions of each concept. If you have a concept like the economy, what words would indicate the economy. That is what you have to figure out. To help with that there are some tools available. There are dictionary building tools like in the program WordStat if you put in a word you can find out a bunch of other things related to that word – antonyms, synonyms, similar terms, hyponyms, homonyms and other word classes – you can actually use dictionary building to help create the features. Also, wild card specifications are important. You have a root from – pleasure, pleasurable, pleasured. You can actually use wild card specifications in a lot of these dictionaries so you can get all forms of the word that might appear.

Another thing that is important is the exclude list. In the Content Analysis Guidebook online we actually have some, we don't have that many dictionaries but we do have a standard exclude list. One thing if you get a frequency list of a bunch of open-ended text what will end up coming out if you don't do anything is you will get a lot of pronouns and simple verbs. An exclude list will get rid of those. They will exclude those so you can actually see the substantive words. Exclude lists are important. There is one available on the website. Most programs do have something like that.

A final thing I wanted to talk about is emergent coding. This is another option that is available. This is when dimensions or patterns of text are derived from the data at hand. You have text without any preset dictionaries, you put them into this program and it gives you results instantly. It is easy; you don't have to create dictionaries. The problem here is that it is less accepted due to the deviation from the positivists, a priori assumption of content analysis and communication. We usually have hypotheses research questions; we don't just put a text in to see what comes out. It leaves a lot to inference. It is a lot more qualitative than I think pure content analysts are comfortable with. But I think it can be useful in

the early stages of the study if you want to just figure out what is in your text, what might be there. In my own work we have used a program called CATPAC. It has been pretty helpful to identify some things that are there. It is beyond just simple frequency counts. And also I think it can get better technologically with advances in artificial intelligence. I will talk about that, it is something that is out there. I don't study this as much as I used to but I am a video games person. I study video games. The games now have really sophisticated artificial intelligence. That is where the money is so that is where they put it. If some of that could be applied to computer text analysis I think some advances could be made. That is emergent coding.

Now what I am going to do is – I gave you a list you can take a look at that. Here are some sample analyses by 3 of my favorite CATA programs, one for each of the types of coding we just discussed. I will show you an example of CATPAC which is an emergent coding program. Then I will show you a program called Diction. It is a program with some pretty good standard dictionaries. And finally WordStat which is a program that has standard dictionaries but really it is more useful for custom dictionaries. It is also tailored to an analysis of open-ended survey data. I was able to go in with that and analyze some of the ANES data. Diction, I think is better for the standard dictionaries.

CATPAC was created by Joe Woelfel at the University of Buffalo. It is a networking program that uses a real network approach, identifying the most frequent words and determining patterns of connection based on co-occurrence. It uses a scanning window. I think someone talked about this kind of approach and questioned it. I think there are valid concerns about this approach. It uses a scanning window to go through and look at 7 words at a time, look at what words go together and it will present cluster analysis results that show co-occurrence. You can put your data in and get results. Here is an example using open-ended text. What we have done with this – you can put in multiple cases in if you have multiple open-ended questions in a survey. You can put those in in case limited form. You get this kind of output – frequency list, alphabetical list. I think what is kind of interesting is that when you go down here, I don't know what this analysis is, but it will give you clusters to show you what words cluster together. That can be kind of helpful but what can really help is if you put it into the other programs that come with CATPAC suite like ThoughtView. This puts your concepts into multidimensional space. You can see what kind of output you can get here. One thing that is kind of cool with this – you can put it into 3-D and put on 3-D glasses and look at this output in 3-D and spin it around. It is kind of freaky. You can see how some of these concepts go together. It is pretty cool. One thing about these – there are questions about the validity of this kind of approach. I don't know if I would ever try to publish studies until this stuff is worked out a little better. But if you are dealing with clients, if you have to show them some results, this kind of thing can be really cool and gives them a sense of what is in the data. It can give you ideas.

Here is an example; I just worked with Kim Neuendorf on a series of studies. We were asked to write a chapter in a book on the concept of identity on how to use content analysis to measure identity. We talk about human coding options and also CATA options for the measurement of identity. We did a small study as part of this. We analyzed open-ended survey responses to a question – describe yourself as a typical American – in CATPAC. We entered the responses, again in case delimited form. Some of these programs have problems with entering data but CATPAC is not that bad. What the results of this study showed was a clustering of what we called practical terms. Again it is inferential, it is not the best science but it is better than nothing. Practical terms vs. other considerations. Here are a couple of examples – here is a 3-D representation. If you look at it in 2-D you can see words like money, goals – those cluster together – tangible, materialistic things perhaps. Over on the other side you have religion, culture, children – family values type things. You can see things going on through a program like CATPAC. It is limited in a sense but it can do some things. One thing that is being worked on too with this – the ability to compare multiple responses – if you have different questions. We had typical American and how do you describe yourself as an American – people are working on programs now that you can statistically compare some of the solutions across multiple questions.

Diction, I will run through real quick. It was created by Roderick Hart out of the University of Texas. He created it mainly for political discourse. It can do a lot more. It has built in standard dictionaries and then 5 master variables that are combinations – activity, optimism, certainty, realism and commonality. You

can analyze text along those dimensions. Custom dictionaries are possible. You can do individual or multiple passages. Here is an example for Diction. It looks like a speech by Bill Clinton, Bob Dole, Richard Nixon, Unabomber text. There are different variables that will come out of this. You will get the frequency with which each of these concepts occurs. Then one of the cool things about this is that you will get a normal range for various types of text. What Roderick Hart did is – he ran a bunch of text with the Diction program and he got normal range for different types of text – thousands in some cases. Then you can see how your text or series of text compares to other text in a particular category or another category. For example, entertainment/TV drama is going to be transcribed, the scripts from those – here are the normal range that these things have on these concepts – you can see how yours compares. Here are some of the options you have for the normal range. If you look at entertainment, journalism, politics, scholarship. I don't think there is anything like open-ended political data or anything but there are some that are close to that. You can compare the normal range and then import your own dictionaries as well. You can put any kind of scores that you get into programs like SPSS.

Here is an example that may help with this from the study I mentioned. We analyzed the first State of the Union address by each of the last 9 presidents. We got the text from an online archive. We put them into a separate file. It was a little bit of work but not too bad. I just wanted to show how Bush's first State of the Union address compares to other political speeches. So there is a normal range for political speeches that was done by putting in a bunch of political speeches. You can see how these compare. Here are all the concepts, if you look on the left side, these are all the concepts. Diction. You get the frequency, how much of the speech was numerical terms, ambivalence, etc. I think for the sake of time here I will skip over to the master scores which are a combination of some of the other ones. Activity, optimism, certainty, realism and commonality. You can see here what this star here is that on one of the variables George Bush in his first State of the Union was outside the normal range in his optimism. He was more optimistic than other political speeches. His first State of the Union came after 9/11 so maybe he was trying to reassure the American people. If you get stuff like that it is an example of Diction.

Finally, WordStat, I will have some sample results from that. The SimStat suite is programmed similar to SPSS. I wanted to mention something about SimStat – it gives inter-coder reliability statistics which SPSS doesn't. If you want to use some of those beyond just percent agreement, look at {?}.

This is an add-on. It is well suited for open-ended data as long as you create your own dictionaries. The standard dictionaries are bad. I looked at the WordStat website today and the author actually has a bunch of dictionaries on there. There is also KWIC.

Here is some sample analyses I did with some stuff that is built in. Basically what you can see here – this is how the data looks. This is a study with 3 variables: gender, age group and then text is an open-ended survey response – it could be but in this case it is a sample survey of analyzing personal ads. You could see that if you click on memo it doesn't show the whole text there but you can click on it to see the whole text. In each case you have a bunch of text for each person. Then you can go through and run analysis based on that. You select independent and dependent variables. Here is the dictionary – here are the things they are looking at in personal ads. I'm going to skip ahead to the thing that I did with the data. Here you can see the frequencies – appearance appeared most frequently in these ads. It is what the people are looking for. You can see how they cluster, finance kind of came out separate. They did comparisons by gender so if you look at women – if you look at communications – women seem to want communication more. Then you can do multidimensional space. You can do a lot of cool output with KWIC.

Real quick, I don't have a lot of time, but I wanted to talk about some ANES. Question – what do you think is the most important problem facing this country? Use this program – here are some sample results. You can see frequency lists – if you looked at the words here – I used the standard dictionary – some of the codes like AE would need to be added to the exclude list. If you work that out – economy, jobs – those things are being mentioned a lot. You can see some other things – homeless, drugs – you can get a sense for what is in there.

KWIC – I used that same personal ad dictionary to look – work – you can see how work is mentioned by people in the study. Here is what I did too – you can output so if you have dictionary categories – some of those personal ads the words appeared in the ANES output so I was able to export that, automatically code it back into SPSS type form. You have a dictionary in here with the economy, homelessness, jobs and if someone mentioned one of those words then that variable will get a score.

Just on strengths and weaknesses: Quick but still not as good as human coding, some validity concerns I have mentioned. Recognition – I have used it at early stages. I think this is fine for exploratory type purposes, data reduction. Think about statistics for using data reduction. This can get a bunch of words down into a smaller number of words. If you have good dictionaries, algorithms with some cross validation too, I think for certain research questions/certain data it can be pretty valuable. It could be used in and of itself. Standard dictionaries, I mentioned trusting only those with evidence of validity and effectiveness.

Price I want to mention too. I think this is a big issue with these. Freeware, I didn't mention any of these. VBPro is pretty good. It is DOS based. It does some of the basic stuff. Relatively inexpensive – student CATPAC and Diction. Expensive – Full CATPAC and WordStat are over $1000. I looked at WordStat and SimPac and you can get them packaged together if you are academic for $700. Academics, most of us don't have a lot of money to work with but this is pretty cool stuff.

This is what I have for the future – making datasets available – when I do statistics classes my students really get into the General Social Surveys because it is available in SPSS. If some of the ANES data was available that would be helpful. Encourage people to use them. The ability needs to go a little longer. Also a combination between machine coding and human coding would be pretty valuable. Also, coding of audio/visual is coming along.

Sorry I went a little over. The website is up if you want to check out some of the programs that way.

Question: Does SPSS also now have a content analysis modular?
Answer: No, actually it was dropped. They had one for awhile. I think it was called TextSmart.
Answer: It has been replaced by Stats which does text analysis. You can have the option of getting that. It doesn't really do anything particularly valuable.

Question: {?}
Answer: If they are measuring a certain concept – how are they measuring it? I think it is important to have some information about that.
Question: {?}
Answer: I would say about half of them do it from what I have seen. The quantitative stuff is pretty rare. I don't think the people really understand it. I think there is also the desire to protect, people don't want their stuff copied or stolen, so they don't want to release everything. They do release the words. Like with Diction, you see all the words they make up. That definitely helps. That gives me a degree of confidence for sure.

Question: You mentioned very quickly in the end about possibly combining the human coding and the machine coding.
Answer: Yes, BCS that I have been looking at is pretty impressive. That presentation is coming up. I can't wait to hear that.
Question: Would it be possible to develop a program that {?}
Answer: Absolutely. I surprised it hasn't been done. It seems like people create these programs – I think for our purposes this seems to be like the ideal solution. I haven't seen a program that emphasizes that except for some of the qualitative ones.
Question: Stuck in there would be the coder would put in or highlight the words of the respondent and then the computer would look up and find several codes that the coder would have immediately.
Answer: That would be incredibly valuable. If anyone here can do that, it would be a great thing.

Question: {?}
Answer: I think I heard some mention of that, some of the program sites that I saw. There is a lot of good stuff online. It is just a matter of finding it.

Question from Skip: There are two things that Jon and I wanted to introduce both to this conversation and some of the next presentations we are going to do. One has to do with the respect to our study, we study politics. Some politics is practical but some of it is ethical and moral and full of values. When politics gets to be about values, meanings get textual. A lot of what people are arguing about is the meaning of certain things, like a free market. What is a free market? It is not just a technical definition, it is how does it relate to other things – those are things that are being tested. A lot of language is used strategically. Because it is a contest among parties, people know that certain words have trigger meanings so meanings of words can be very purposely different by different speakers for different audiences. I want to put that on the table as we think about computer programs and the challenges that might introduce. We talked before – does everyone agree what violence is – that is an interesting question. In politics you not only have that but you have potential of people constantly trying to change the meaning of that word as part of their strategy for gaining votes. That is one thing we wanted to put on the table. The other is a more practical thing that affects the surveys in general. Suppose we were to go down the road of computer coding, and we are going to see a lot more of this, it sounds like maybe some of the more powerful and value adding utilities are private, at least some of them. If we were to work with a proprietary firm it seems like by definition you would be working with a firm that had proprietary ideas. They have this software that nobody else does. It would be the basis for paying them. We are curious about what you think about that. It seems like the advantage is that you get the increased power of working with these utilities, the cost is you sacrifice transparency. You don't have the code that allows that firm to derive revenue. We can talk about that now or we can introduce it later.
Answer: It strikes me that the comment that Randy Olsen made – if you release this {?}

Question: I think as we go forward we really need to have a plan. Plan A might be – what if we are permitted to {?}. Plan B would be what if we are not. We are talking amongst us and very fine arguments are being made so I am just putting it out there. We are not sure if it is feasible. We need a plan, I think, either way.
Answer: To address the second issue there are open source applications that are power based {?}