Open-ended Questions and Text Analysis
Roel Popping

Just like Klaus I have to tell something about doing text analysis as part of a survey. My emphasis is different from his, so there will hardly be overlap. He didn't talk about the topics I will talk about, for example reliability. If one wants to talk about text analysis, there are many topics to discuss. Actually there is far too much to say so please don't be surprised when I skip some sheets. This because I realize we will have some problems because of the 30 minutes limit. I will try to tell what I believe is important.

There are three topics I would like to talk about: open-ended questions, a little bit about text analysis and here focus especially on the steps one has to take to apply text analysis in the research. In my view inter-coder reliability is extremely important, so in the third place I should say something about that. If you don't use it, it is really a shame.

If one has open-ended questions in a study, as was said this morning, one can distinguish several types of questions. My classification is close to what we had this morning. At the bottom of the list on the screen is a reference to the situation in which the investigator has some fixed answers for his question. The set of categories should be exhaustive and therefore the investigator has introduced the possibility of "other" and he wants to make the respondents have the opportunity to tell what this "other" means. He hopes that no one will use this facility, but to be complete he must offer it. The type in the middle is what I call the apparent open-ended question. This type is at hand when it is impossible to list all answering possibilities. An example is the occupation. In case the occupation is asked for, a knowledge base that is hanging under the questionnaire or that is added afterwards can look for the exact ISCO-code that corresponds to the occupation entered. In earlier research using a knowledge base where a retest was performed we found a reliability of about 87%. The differences were caused by the other information provided by the respondents the second time. For example not the actual situation of being a pensioner was mentioned, but the last performed profession: a carpenter. The main thing to talk about is the general open-ended questions. I really want to talk about that.

When talking about open-ended questions in surveys, first comes of course: why do you want the open-ended questions? It was already said that there are all kinds of arguments that have to do with this. One can look at validity. One can look at it as it is very satisfactory for the respondent. One can look at the possibility of getting unforeseen answers that might be fairly relevant in some fields. One can look at the possibility of control – one knows that the respondents did understand what one was asking. If one wants to use the open-ended question then most important is that the proper question is posed. One should, for example, not ask "what are important problems," but one should ask "what are the important political problems today in your country." If one is not that specific one will get all kinds of answers that one does not want to have and then of course one might not get what one wants.
With respect to the analysis I already mention that sometimes there are persons who do not give an answer to the open-ended questions. Please do not automatically say that this is a missing answer. It might be that it was just not appropriate for the person who was interviewed to give an answer. This is different from that it was not asked at all. There are studies that really deal with this problem. The problem that there is a reason why the people do not have an answer. Do not say automatically that it is something different.

If the investigator has the correct question then he can move to the text analysis. In the text analysis the condition is that one has a critical question. If the investigator doesn't have that he will already have a problem. The input for the text analysis is always the texts as the investigator has them. Be careful those texts, answers to open-ended questions, are not always meaningful and interpretable. They may consist of catchwords or cryptic words and contain a lot of grammatical spelling errors. No matter how the analysis is performed, but especially when it is done by the computer, the investigator should be able to handle all those errors.

The output that one will get from the text analysis is always a matrix of occurrences of themes. By that I mean that you have in the row the respondents, to be recognized by an identification number, in the columns are the themes. What you get in the cells is the number of occurrences of the corresponding

theme as mentioned by the respondent in the answer to the open-ended question. The occurrences might be marked once or they might be there several times. This just depends on what you are interested in.

This is basically the kind of output you get. There are other situations where it is more complex. But, this is the very basic idea. From here one can start with the text analysis. And from here comes the moment that one has to make choices. If an investigator starts text analysis he can have three things in mind for the text analysis that have to do with the approach one is following. Most people doing text analysis are doing what we call the thematic approach. They are interested in the occurrence of certain themes or concepts. That is what one wants to know. But it could also be that one is interested in a semantic approach. There is a text, the basis of the text is a sentence, the sentence usually consists of one or more clauses and a clause consists of a subject, a verb and an object. One can look at the relation between these three. One can even have the following: the subject in one sentence is the object in another sentence. Based on this networks can be constructed. These networks can be analyzed for example by looking at the meaning of chains of lines, the verbs in the original clauses.

These are three approaches; with regard to each different ways of coding can be applied. These are the instrumental and the representational way of coding. If the instrumental way of coding is followed the perspective of the investigator is followed. If machine coding is performed, one always follows the perspective of the investigator. This way of coding is quite easy to perform. An interesting fact is that one is usually interested in the person who is sending out the message. What does that person want to say? What is the message? If you are interested in that information then the representational way of coding must be used. To give an example, at this moment I am coding data from the communist countries in Europe before 1989. We have to read between the lines to find what people mean in the text. They present a very nice story. It is perfect, but between the lines you can read that in several cases the author wants to tell some other things. If the representational way of coding is applied you get the opportunity to code this other story.

I do not want to talk about semantic and network approaches. The speaker after me this afternoon will mention the semantic approach. Tomorrow we will hear about the network approach. So it is coming anyway.

To start with, as an investigator one has to decide which approach to coding will be followed. In case the instrumental way of coding is followed the coding can be performed by a machine. It can also be done by hand. If the representational way is followed you can only do it by hand, at least in my view. Of course the computer can be used as a management tool and please use it as such, otherwise you will get lost in your data. The computer is very helpful here. The computer does not do the coding; it just remembers what you have been doing.

Most people start with code sheets. The code sheet contains a specific question, often contains themes and sometimes there are remarks on how to look at things. The task of the coder is to find that part of the text that refers to a specific question or a specific theme. If that piece of text is found then the category can be assigned. It is also possible that a piece of text one is hoping for is not found. Do not be surprised if this occurs, such a piece does not always have to be there.

When the instrumental way is applied the investigator can use machine coding and that is what most people do these days. When they use machine coding they start with a search entry. The search entry is used to find a piece of text (a word or a phrase) in the text that is to be analyzed. Via the search entry that piece of text is linked to the corresponding theme and in this way it denotes one more occurrence of the theme. This is the basic idea. If this approach is followed you can handle long texts. In general coding goes quickly. Of course, it can take hours, because the text is so long. But nevertheless it goes fairly fast.

In order to perform a text analysis one should have a dictionary available. In making a dictionary one will run into a lot of problems, however, and answers to these problems must be found. To indicate some problems, I will start from the bottom of my list.

One can have words indicating judgments. Is it possible to have the computer recognize words in the text indicating a judgment. More important is how to use the judgment, because it is related to something else: often the subject or the object, for example: "it is a good attitude."

In the text there are often pronouns – he/she – that kind of things – they refer to Bill or George. But maybe the person himself is not interesting. Maybe it is relevant that a reference is made to the role of president or to the other role Bill or George has at that moment. It is up to the investigator to find a way to make sure that these pronouns are coded in the right way. There are several ways to do it. One is to add these roles in the original text.

The text might contain negations. The computer program has to recognize these. Some computer programs claim that they do so, they have the possibility to distinguish between "I am going" and "I am not going", but sometimes I have my doubts about this. It is possible of course to say the problem can be solved. One can say that "I am not going" should be replaced by "I am staying." Or one says that if "it is not beautiful" then "it is ugly." Here however, one should be careful in doing this because it is possible one gives another meaning to the text and maybe an unintended meaning.

Finally, there are the ambiguous words or remarks. There are three kinds of ambiguity. The idiomatic ambiguity refers to the situation in which it seems a clear expression is presented at the surface, but actually something else is meant. An example is: "You will eat your words." This really does not mean that you will eat them, but that you will regret what you were saying. If there is a teenager saying that yesterday evening she has been to an awful movie, you must know that she usually means it is a very beautiful movie. An investigator must be familiar with this kind of terminology. There might be a restriction however. In my country an awful movie is a very beautiful movie. It does not have to be true in other countries. Next, there is the illocutionary ambiguity. This characterizes statements with meanings that vary as a function of statements made prior to them in context. If you are talking about "Joe bought that camera," it might be something very stupid, an awful thing. On the other hand, it might be very, very good. Given that you know what happened before, you know which answer applies. Also if you are talking about "she is a doctor," what do you mean then by a doctor? From the context you might get that information but if you just read the sentence you will not get that. I skip the third type, the relevance ambiguity as this only applies to networks.

It was said before that one needs a dictionary. It incorporates a theoretically motivated classification scheme for the words or word combinations occurring in a text under study. It consists of themes, search entries (per theme) and rules for how to proceed in case not only ambiguity is encountered, but also if one of the other issues just mentioned is found. I mentioned four issues, there are more.

If a dictionary is constructed, it will look something like the following. The theme might be the self reference. And the self reference has a number of search entries, words or phrases to be found in the text indicating the self reference (I, we, us). When constructing the dictionary, you have to find these search entries. The role of the investigator is to decide about the themes to be used and about the search entries that are linked to each theme. There are questions with regard to finding the themes, but I will skip that part in the presentation when I come there due to lack of time and because it is quite open. So it is not that informative. It is usually something that you have to find out for yourself.

If the representational way of coding is applied then it is the coder who has to find the text fragments and tell whether a certain theme is occurring in the fragment – yes or no. It means that the coder is the one who is giving the interpretation. The method allows the coder to read between the lines, which is a strong point. One has the opportunity to detect all kinds of information that was unintended but that is nonetheless available or one finds information that officially could not be said but is available. There are problems with the approach. First, it is really time consuming and the other problem is that coders have to know how to do their job. For this one might need a coder training and one has to look at reliability, which implies that at least part of the coding task must be performed twice.

In order to show that both methods can result in different findings an example is presented. The example comes from a theory saying "I (self) am fair" and "you (other) are unfair." The investigator is looking at the co-occurrences of self and fair and of other and unfair. Data have been coded following both the instrumental and the representational view. If you look at the data on the screen you see there are really differences in the outcome when these ways of coding have been applied. As an investigator you have to be aware of this fact. It may help you in deciding on the method to use.

Now comes the part that I will skip because I said that I wish to talk about reliability. I have been telling that there are a number of steps, that you have a number of decisions to make and that these will determine the kind of data that you get. I belong to the people who really believe that hand coding is most valuable, so I do that. The consequence is that I have to do a reliability check. Here one can talk about stability. Here the question is just when a coder does a job and then does the job a second time, will he come to the same results? If so, it means he himself is stable. Accuracy is that you can compare what you are doing to a standard. In our field that will be very difficult, but in medical settings it is extremely important to have such a comparison, because the decision (medical diagnosis) must be correct. Finally there is agreement. People very often talk about agreement as a kind of association. That really is not correct. Agreement is not association. If you have association you can predict one variable from another variable and in the case of agreement that is not what you do. In the case of agreement it is important that you decide about the extent to which different coders come to the same conclusion and assign the same category. When you are talking about agreement this is the thing you are talking about. The coders should come to the same result. If you are talking about this skills become important. I will come to that later.

The computation of agreement, reliability is basically based on the proportion of observations on which the coders agree. But there is an amount of agreement that is found on the basis of chance alone. It is very good to control for this amount of change expected agreement in the index. If there are two coders who each use two categories indicating something takes place or does not take place, the assignments can be compared within a cross tabulation. Say cell A is the diagonal cell containing the observations on which both coders agree the something took place, and cell D is the diagonal cell containing the observations where the something did not take place according to the coders. Cells B and C contain the observations where the something took place according to one coder and not according to the other one. Instead of number of observations it is also possible to present proportions in the cells.

We can now look at the amount of observed agreement – that is what we are interested in. This is the proportion of observations in cells A and D. We can also look at the disagreements. We would be very happy if cells B and C are empty. That would mean there is complete agreement. The next thing to talk about is the agreement that is expected by chance. We have two coders. They performed the codings in a situation where we do not know what exactly is the perfect coding. We can talk about the number of times they used each of the categories, but as there is no reason to assume one coder did a more correct job than the other one, it is best to talk about the distribution averaged over all the coders. In the cells on the diagonal you will now get the average proportion of times the corresponding category was used by one coder times the average proportion of times it was used by the other coder. These proportions are for each category the same for each coder. Taking the proportions in the diagonal cells together gives the expected agreement. The final index contain in the numerator the difference between observed an expected proportion of agreement, and in the denominator the difference between the maximum amount of agreement, the proportion 1 by definition, and the expected proportion.

Of course, this idea can be extended, because it is not always so that there are two coders using two categories. The index can be extended to the number of codings you have per unit or per observation. This can be different. It can sometimes be that the observation is coded by two people at a time, or three or four people – that is all possible. There certainly can be more than two categories. If there are more than 2 categories it might be that the categories are ordered or even on an interval level of measurement. One might be interested in agreement in a fixed category because it tells you about where exactly the problems are or whether there are no specific problems at all. One can even talk about the definition of agreement. One can compare all the pairs of coders. But one can also say that it is only ok when all coders agree in assigning the observation to the same category. If one coder is not agreeing then it is wrong. Or maybe the investigator is already satisfied when k out of the m coders have the same opinion. However, please don't do this. Only look at the pairwise agreement. In an article that will be published next year I will explain why. If pairwise agreement is considered, you get matched formulas and you can compute all the specific forms of the index that you need. There are other opinions with regard to the computation of agreement. What I just told, is only one way to do it.

Coders, they are really important people. They should be aware of several things – the goal of the investigation; they should know the meaning of the categories used; they should be trained in identifying the target behavior one is looking for; if they are in the position to look at several categories at once they should know which one is most important; and if there is a preference they might have the skills to judge this.

There are a number of steps to be followed to come to this kind of decision. Carl, the next speaker, and I performed a study in which modality in clauses and rationales play a role and in which we look at the reliability of the assignments. When one has to do this first one has to find out whether the text contains a modal clause. Second, when it contains a modal clause, do the coders identify the same one. If so, do they agree on the identity of that clause? Only then one can talk about the clause's rationale. The type of clause and the rationale are two issues that were relevant in the study and of course we hoped to find a high coder agreement for the two. The decision just mentioned show that there are some steps that also must be taken to come to the agreement. It means that there is far more than one might be thinking of when first confronted with the coding and that this all is more than is expressed in the outcome of the index.

So then comes the coder training. Actually this consists of two parts. The first part is that each coder has to assign the codes, the modal clauses and the rationales. The coders have to develop a set of coding rules. The reproducibility of the assignments is very important. We wrote a complete article about the relation between the modalities and the rationales, but we also wrote a complete article only about the coding rules. These are so complex and it is so important that people are informed about it. With respect to reproducibility it is important that the rules followed are explicit. They should be known somewhere.

I have very quickly talked about these three things, open-ended questions, text analysis, where I mentioned a number of decisions and I have talked about reliability. What I haven't mentioned is that when you are computing reliability you also have to decide on when this reliability is good enough. There are a number of people who say it is ok when it is over .80 or .70. Please do not do that. Think about the coding task that is to be performed, about the complexity of the task and take these into account when you make your decision. Do not follow a rule that someone has mentioned because it sounds very nice.

Then there is the training of course. This is very complicated. One can talk about how to organize it – do the coders have to use the data or a part of the data that was used and how many coders have to perform the task? How does the training have to be developed? What to pay attention to during the training. How much time do coders need to learn? In the medical field people are doing a lot of research on students following the physician. The assignments by the students are compared to those by the physician, the expert. IN the first years of the study, the percent of agreement is about .40. This implies that if you have the index that I mentioned before it is even lower. The agreement (the quality of the assignment) is increasing, of course, during the study but still it is not going very quickly.

One needs a lot of energy to get people at that level. If you have a good reliability, that is my final remark, then be aware it might be that the results that the coders did very well is overestimated. This because the coders have been telling each other, in an informal way, on how to proceed in some instances. Be aware of this possibility. It can come this far because the coding process is not transparent and therefore no one knows how it works. That is not useful. Be very clear about how things have been performed.