Analyzing open-ended questions by means of text analysis procedures

Roel Popping
Department of Sociology
University of Groningen

Paper Conference on Optimal Coding of Open-Ended Survey Data, to be held December 4-5, 2008, in Ann Arbor, Michigan, at the University of Michigan.

First draft, not corrected. Do not cite.

Abstract

Assume one has open-ended questions in a survey and seriously wants to analyse the answers to these questions. Now text analysis might be applied. This talk discusses a number of choices to be made when a thematic text analysis is to be applied. It starts with requirements to be posed to the open-ended questions themselves and sketches choices in the development of a category system. Here the coding comes immediately into view. Coding can be performed from an instrumental or a representational perspective. In the first the coding is performed from the point of view of the investigator, it can be performed in a run of a computer program. In the second the point of view of the respondent is acknowledged. Now the computer can be used as a management tool, but the coding itself must be performed by a human coder. The choice for one of these methods depends on what the investigator is looking for and has consequences for the way how to proceed. When the representational way of coding is applied also questions about intercoder reliability must be posed.

Introduction

An open-ended question in a survey or public opinion poll is an unstructured question in which possible answers are not suggested, and the respondent answers it in his or her own words. The question is framed in such a way as to encourage the explanation of the answers and reactions to the question with a sentence, a paragraph, or even a page or more, depending on the survey. It should be a full expression of an opinion, containing nuances, rather instead of merely having to select an answer from a predetermined set of response categories. This allows investigators to better access the respondents' true feelings on an issue. Open-ended questions tend to be more objective and less leading than closed-ended questions.

Open-ended questions usually begin with a how, what, when, where, and why (such as "What are in your view the most important political problems in your country today?") and provide qualitative instead of quantitative information.

Open-ended questions have a great diversity of responses. This is certainly true when one keeps in mind that not all respondents have the same ability of expressing or the same style of writing. One might be afraid respondents are not able to get their opinion on paper and that they only articulate a response. Geer (1988) has shown they are able to formulate their attitudes. Certainly when sensitive information is to be elicited, such as information about sexual assault or drug usage, open-ended questions seem to be better for obtaining information. Another benefit is that these questions cut down on two types of response error. Respondents are not likely to forget the answers

they have to choose from if they are given the chance to respond freely and open-ended questions simply do not allow respondents to disregard reading the questions and just "fill in" the survey with all the same answers. Therefore the answer is very reliable.

It looks as if open-ended questions are a nice alternative to closed ones in case the investigator beliefs there is more to be answered than can be captured in a question with fixed response possibilities.

Some disadvantages might also be mentioned. The reliability of the answers to open-ended questions might in reverse to what was mentioned before be low, as possibly the respondent will only report what comes to the mind and that might be a part of what could have been covered when closed questions had been used. Another problem might be that even with respect to one issue, one respondent might produce far more information than another respondent. A consequence is that the investigator must be convinced about the surplus value of an open-ended question.

The investigator will finally run into another difficulty: the answers must be analyzed. Text analysis seems the tool to be used now. This text starts with a number of remarks about open-ended questions to be used in a survey. Next the opening to text analysis is made. A number of possibilities is shown, but these also contain some restrictions. The text analysis might be based on human judgments. Now intercoder agreement comes up, as one needs a reliability check. This all should give insight in how to use open-ended questions in a research project so that the data obtained are really used.

Open-ended questions

Three types of open-ended questions are distinguished:

1. The general open-ended question;
2. Apparent open-ended questions;
3. Addition to the category “other”.

To start with the last mentioned type, a set of categories belonging to a question should be exhaustive. In order to realize this often the category “other” is used as the last one. As this category is not informative the respondent gets the opportunity to specify what is in the actual situation meant by “other”. The respondent does not have the bad feeling that his or her opinion can not be expressed. The investigator might get an answer that is after all very valuable. In practice however, the investigator hopes that the respondents will not use this category. An example from an election study is (Van der Eijk et al, 1988: 192):

> “Assume people would be obliged to go and vote, which party you would be voting for now.”

The last one of the response categories is:

> “Other, namely …..”

Here the respondent is supposed to explain what he has in mind and what is not included in the previously mentioned categories.

Sometimes it is impossible to list all answering possibilities; therefore the question is presented as if it were an open-ended question. Examples of such questions are the ones for the year of birth or for the occupation. The year or label of the occupation is registered. 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 entered occupation. For some occupations it is necessary to know more than just the occupation, also information might be needed on the being self-

employed or not, or on the number of people supervised. Popping (1995, 1997) showed that the computer does this registration more accurate than a human coder (who is familiar with the ISCO-classification) does or than a person working at a check in desk.

When investigators talk about open-ended questions they usually have questions in mind that result in stories by the respondents. These are the general open-ended questions. Examples from election studies are (Van der Eijk et al, 1988):

"Why did you vote for this party" (p. 139, 260)

"Are there things you appreciate very much in this party or for which you have sympathy" (p. 162)

"Can you tell how you understand 'left' in politics" (p. 24, 167)

Sometimes the question consists of several sub questions or has the form of an incomplete utterance to be completed by respondent.

The answers to these questions will usually be descriptive. People will present lists, will tell what they feel or think what their intentions are. Sometimes it will be mentioned explicitly, sometimes one has to read it between the lines.

Open-ended questions must be specific to provide meaningful, interpretable data; therefore the formulation of the question is relevant. This is especially true when there is no interviewer, who can give a reaction to the answer. It is not possible to ask for more detail or to send the respondent into another direction. The question must include a direction. The answer must be that good that one can use it. Compare the differences in the following examples. The first question leaves the interpretation up to the respondents, which could differ from the question's intended meaning:

"What are important problems?"

The second question, which was already mentioned, provides specific information on what is being gathered:

"What are the important political problems in your country today?"

The answer written down by the respondent might be formulated in perfect sentences, but it is also possible it consists of catchwords or cryptic words and contains a lot of grammatical spelling mistakes. During the analysis this is to be taken into account.

Here is the first decision moment for the investigator: is the open-ended question well formulated, so that we get the information as intended.

Now one might have additional information, one could not get by using closed questions, but this information still needs to be classified in some way. For this text analysis is a proper tool.

Few examples are known in which investigators performed a (computerized) text analysis on the answers to open-ended questions. McDonald (1982) asked respondents in an open-ended question to comment on a product. The answers were entered into a text file, and were next classified by using the computer. In this way McDonald could get a more detailed judgment of the way the product was advertised in the media than if he would have asked the respondents.

At this moment already one remark can be made with respect to the analysis when the results of the text analysis are linked to the other data that were collected. When no answer is given to an open-ended question, this should not automatically be coded as a missing value. The respondent might have considered the question as a peripheral topic, which does not have to be discussed.

Text analysis

In this text two restrictions are followed. First I will only focus on text analysis that is supported by a computer program. Before three types of computer supported text analysis have been distinguished (Roberts & Popping, 1993). These are thematic text analysis (in which one looks for the occurrence or co-occurrence of themes), the semantic text analysis (where one uses the subject – verb – object [SVO] relations as found in sentences), and the network text analysis (where the SVO relations are even combined into networks). My expectation is that in the situation of analyzing answers to open-ended questions most of all thematic text analysis applies. Because of this and for reasons of space available the second restriction is that I only consider this type of analysis.

The input for the text analysis consists of as many pieces of texts as there are respondents in the study. It is possible that a piece of text is empty when the respondent did not give an answer. Attached to the piece of text should be some identification mark. The output consists of a matrix with in the row the texts (which might even be split up) and in the columns the themes. The cells inform about the number of times the corresponding theme is found in the text that is considered or just whether it is found. One exception will be mentioned later on.

A computer-program counts the occurrence of themes in the texts, here the answers by the respondents. In general these are indicated as units of analysis or sampling units. In text analysis studies a distinction is made between units of analysis and recording units. In the situation of analyzing open-ended questions the unit of analysis would refer to the respondent. Recording units are "the specific segment of content that is characterized by placing it in a given category" (Holsti, 1969: 116). A unit of analysis can consist of several recording units. Therefore it might be necessary to bring the codes at the very end to the level of the unit of analysis. How this is done depends on the actual research situation. In text analysis studies often weighting is used.

The coding tasks, which might be performed by human coders or by a computer program, are divided into three types (Crittenden & Hill, 1971: 1078): A, B1, and B2. "Type A coding tasks require a coder to find a specific answer to an explicit question at a given place on an instrument. Type B1 coding tasks involve locating relevant information within a larger context ..., type B2 coding tasks are those where the coder has not only to locate relevant information, but also to evaluate the relative importance of two or more possible responses to arrive at a single code" (Montgomery & Crittenden, 1977: 236). This distinction will come back later.

An important question in text analysis is from whose point of view the data are to be analyzed. In the instrumental approach texts are interpreted according to the researcher's theory. The approach ignores the meanings that the texts' authors may have intended. When the representational perspective is applied, texts are used as a means to understand the author's meaning (Shapiro, 1997).

Today many investigators still follow the instrumental approach, as is shown by Shapiro. Here a computer program takes care of the coding. In this situation the investigator uses a "fixed dictionary" of thematic categories with a one-to-one correspondence with words and phrases in texts. Researchers using the representational approach must develop (ad hoc) dictionaries that contain themes that reflect the perspectives of the texts' authors. Coders must use sympathetic "understanding" (or "Verstehen") to encode the texts according to the meanings their sources intended. At issue is no longer "how" to encode text (instrumental approach),

but “whether” one chooses to apply one's own theory or one's sources' theories to the texts under analysis. Sometimes one also needs human coders when the instrumental approach is followed.

According to Heinrich (1996) the fundamental advantage of computer-assisted text analysis is the intersubjectivity of the result. Personal evaluations by the coder have no effect on the results. In his study Heinrich used the instrumental approach and a computer program performed the coding task.

Here is the second decision moment for the investigator, should the instrumental or the representational approach be followed.

Instrumentational approach

When the instrumentational approach is followed two ways of coding might be distinguished. In the first the computer is especially used as a management tool, the coding is actually performed by hand. In the second the computer program takes care of the coding.

The investigator has to choose whether hand or machine coding will be performed when the instrumental approach is followed.

The first way, hand coding, is found when it is not possible to automate the coding process. Riffe et al (2005: 17-118) present a so-called codesheet which contains a number of questions and their response categories. Several sets of categories contain some ordering or are such that a computer program can not recognize them. A question is for instance:

“Primary story source”

with categories “debate”, “speech”, “document”, “poll”, “interview”, etc. In the coding process the investigator’s view is followed but human coders are needed to choose the answer that fits best. In most situations one is confronted with a type B1 coding task here. Coding here resembles the coding as is performed when the representational view is followed. Therefore the way will be discussed as part of that approach.

When the second way is followed, a computer program performs the coding task. The investigator uses search entries, these are words or phrases that appear in a text and that refer to a certain theme. For this a dictionary is needed, consisting of a (theoretically motivated) list of themes to be looked for, for each theme the words referring to this theme, and a series of rules telling how to deal in case some difficulty rises. The finding of themes will be addressed later on. Now the computer program does the analysis by looking for search entries in the texts and when found a new occurrence of the corresponding theme is registered. Here one does not need human coders. Advantages of this way of working are that long texts can be handled and that coding goes very fast (once the dictionary is available). However, the investigator has to anticipate on a number of potential problems in advance. For this the rules are needed. Texts contain ambiguous words and negations. How to deal with pronouns (he, his, who), should they be replaced by concrete names or roles. A search entry might refer to two different themes. I will not discuss the problem of formulating search entries, but it is possible a search entry is incorporated in another one. A text might contain typing errors. In the type of text analysis we are considering now a 1:1 correspondence between search entry and text is needed. So, in case of a typing error no occurrence is registered.[1] In reverse, a text might contain a saying specific for a

certain society that has another meaning than what is literally said. Here an occurrence will be registered, while actually there is none. This refers to ambiguity in texts.

As far as ambiguity is concerned, a word might have different meanings, which becomes clear when the context is known. An example is the word “spring”. This word might refer to a season in the year, but also to the beginning of a river. Special types of ambiguity are (Roberts & Popping, 1996):

- Idiomatic ambiguity - characterizes expressions that lack clear meaning for those who are "outsiders" to a particular social group. Teenager's references to "awesome clothes" or "an awesome movie" may have no clear meaning for older people.
- Illocutionary ambiguity - characterizes statements with meanings that vary as a function of statements made prior to them in context. Take the sentence: "She is a doctor.” Is she altruistic, or an achiever?
- (Relevance ambiguity – in network grammars, not relevant in this text.)

The two types of “spring” can be distinguished in the texts by adding some sign to the word (disambiguation), so one will have “spring#1” and “spring#2”. But expressions of the two types of ambiguity mentioned can only be made clear in a computer supported instrumental analysis when they are replaced by an expression that is clear. And now the question becomes whether this is allowed, one might also make unintended changes.

Something similar holds in the situation of negations. Is it allowed to replace “not going” (hopefully the words are immediately following each other) by “staying”? Usually this is solved in a very ad hoc way.

When machine coding is applied the investigator must decide at a given moment on whether the dictionary is complete, on whether ambiguity and negations have been taken care of.

Representational approach

When the representational way of coding is followed the coder selects a text fragment and assigned a theme to this fragment. This way of working allows interpretation by the human coder; the coder can capture the latent meaning of a text. The method is time consuming, as coders have to perform the coding task. Often it is a type B2 coding task. The coder has to read between the lines, not only the manifest content might be looked for, but also he latent meaning might be considered. This meaning can usually one be understood when the point of view of the sender of the message is taken into account. Here is the main difference compared to the instrumental approach to coding.

When this approach is followed, one will be confronted with the question whether the coder did a good job. So a coder training to optimize the coding task, might be, usually will be, necessary. It will be beneficial to have at least a part of the coding task be performed by several coders. This allows computing intercoder reliability: do the codings resemble each other. Here the computer program is most of all a management tool.

The decision moment concerns especially the themes that will be used and how to add themes. But also a decision must be taken on manifest and/or latent themes. The next decision concerns the rules that are to be followed: does the coder know how to recognize a theme, especially a latent one (and can he)? So, is training necessary.

Finally the investigator must decide on intercoder reliability. Is it to be computed and if so: how and when?

Example on results
Table 1 contains data from a study in which 40 sportsmen say something about themselves or their team (Self reference) or about the opponent or the other team (Other reference). What they say has to do with being fair (Fair) or unfair (Unfair). The goal was to test that the Self reference goes together with Fair and the Other reference with Unfair. The data have been coded in twice, once by using the instrumental and once by using the representational approach. 89 recording units have been used.

Table 1 - Coding of four themes using two methods

	Instrumental				*Representational*			
freq	Self	Other	Fair	Unfair	Self	Other	Fair	Unfair
0	61	60	83	63	61	59	83	66
1	24	25	6	25	24	29	6	22
2	4	3	-	1	4	1	-	1
3	-	1	-	-	-	-	-	-

Source: Popping, 2000, data are presented for recording units

Both methods yield different outcomes for the themes Other reference and Unfair. One would like to work with two dichotomous variables (Self or Other reference versus Fair or Unfair behaviour), so new variables might be created. These variables will have four categories, as none reference mentioned and both references mentioned are also possible outcomes. The same is possible with respect to fair or unfair.

The resulting cross-classifications follow in Table 2.

Table 2 - Cross-classifications of reference against fair/unfair

	Instrumental			*Representational*		
	-	Fair	Unfair	-	Fair	Unfair
-	21	2	9	26	1	4
Self	22	4	2	23	5	0
Other	14	0	15	11	0	19

These results show that the due to the approach followed during coding different results are found. Looking at the relevant categories the relation between the variables turns out to be perfect when the representational approach to coding is used, but when the instrumental approach is used two units are placed in a undesired cell of the table. Note that the number of observations differs for the two approaches.

Categories
For the development of theme categories for text analysis in quantitative research two "ideal" types may be distinguished:

- A set of theme categories is developed a priori, based on theory underpinning the research project. These categories are an operationalization of theoretical notions or have been used before;

- The theme categories are “data driven”, i.e. they are constructed a posteriori, based on words or phrases in the texts that are analyzed. It is possible they were developed in the exploratory phase of a project.

Combinations of both types are found frequently. The investigator starts with theoretically derived themes, and adds data-driven ones.

Here too is a decision moment for the investigator, a priori or data driven categories.

In case a variable, together with its set of categories is a priori available, each separate category is in the analysis considered as a theme. The occurrence of the themes is coded. This will go perfectly as long as it will not be that two themes belonging one category set are both coded. These themes should exclude each other. This might be difficult as the text unit often contains all kinds of information. Both political and economic facts might be discussed. Here one should choose for the theme that gets most attention, eventually from some point of view.

The other position is that during the performance of the coding task new themes come up. These themes do not refer (yet) to a more general variable. So here one can expect that several themes are assigned to a text. Deriving themes might start from the frequency of occurrence of words. But, this is not sufficient, one should also consider why and how certain words or phrases occur, especially in relation to the research question at hand. In this consideration process resulting in themes the investigator might rely on qualitative research, because it is more explicit on how codes are developed. Besides corresponding computer programs contain more useful tools. In this process the investigator might also use indicative schemes that have been developed.

At the end the investigator might want to transpose themes into categories of a variable. Here are two possibilities. Themes might be considered as categories of a variable. My themes ‘Self’ and ‘Other reference’ might be categories of a variable ‘Reference’. But actually there are four categories: ‘Self’, ‘Other’, ‘Both Self and Other’, and finally ‘None’. It depends on the goal of the investigation whether the third category can occur. Practice is that a sentence like: “We were impressed by the other party” is realistic.

The second possibility uses characteristics themes have in common. Some computer programs allow performing some kind of cluster analysis, principal component analysis or multidimensional scaling approach on the data matrix in order to find useful dimensions that might represent variables.

Always the investigator decides which themes will be used. In case the instrumental approach is followed, search entries are specified for each of the themes. These search entries will be found in the texts or can be copied from earlier research. It was already mentioned that a search entry can be used for more than one theme. It is up to the investigator to decide how a search entry will be used.

Anyway here is a decision moment: Does the investigator want to use variables having a set of categories, and if so, how to get there?

Often investigators are not that informative on how they did get their search entries. Carey et al (1996) are a good exception. They created the code list completely from the data. They do not present details on how they did it.

Actually unnecessary to say that before a text analysis can begin, the texts need to be preserved in a form that can be analyzed by the computer program that will be used.

Components of categories

Contas (1992: 256) discussed components of categories according to three procedural elements: 1) origination, 2) verification, and 3) nomination.

Origination refers to the locus of category construction: "Where does the responsibility or authority for the creation of categories reside?" Contas distinguishes five places: participants (as opposed to investigator – I interpret this as to whether the representational or instrumental approach must be followed), programs (categories are sets of goals or objectives), investigative (personal interests, investigative perspective – intellectual constructions of the investigator), literature (based on statements or conclusion from literature), interpretative (based on data).

Verification details the strategies used to support the creation and application of categories: "On what grounds can one justify the creation or existence of a given set of categories?" Here Contas distinguishes sources of justification: external (panel of experts), rational (relies on logic and reasoning), referential (uses existing research findings or theoretical arguments to justify categories), empirical (examines the coverage and distinctiveness reflected by categories), technical (borrows procedures from quantitative methods to answer verification questions) and participative (participants must get the opportunity to modify codings).

Nomination, finally, is concerned with the naming of categories: "What is the source of a name used to identify a given category?" Here the same sources are met as under categorization.

Apart from using such questions investigators might start with general phenomena that are not content specific but point to general domains. One such scheme is presented by Lofland (1971: 14–5). The phenomena are sorted from microscopic to macroscopic:

1. Acts. Action in a situation that is temporally brief, consuming only a few seconds, minutes or hours.
2. Activities. Action in a setting of more major duration – days, weeks, months – constituting significant elements of persons' involvements.
3. Meanings. The verbal productions of participants that define and direct action.
4. Participation. Persons' holistic involvement in or adaptation to a situation or setting under study.
5. Relationships. Interrelationships among several persons considered simultaneously.
6. Settings. The entire setting under study conceived as the unit of analysis.

Using the schemes by Lofland and Contas, one can now ask:

- What are the characteristics of acts, activities, meanings, participation, relationships, and settings?
- Which forms do they assume?
- Which variations do they display?
- From which perspective should the corresponding theme categories be created?
- What justifies these categories?
- From which perspective is the name of the category derived?

This gives a handle to start developing theme categories.

In qualitative research, theme categories are developed as part of theory construction. The explorative part of a quantitative study might be performed using the tools and computer programs developed for qualitative research to develop the theme categories (and search entries) to be used in the final project. The facility to write memos is always

useful. Programs for text analysis that allow representative coding should contain such a facility, for during this coding process many decisions are made that need to be remembered.

As an example of how the construction of categories is often performed I use one example. Kurasaki (2000) described how she developed a priori her codebook using a grounded theory approach. The steps followed were: annotate five transcripts together with respect to the interview's contents, sort the annotations into similar categories and subcategories and label the thematic categories. If necessary repeat the process and refine the categories. Popping (1992a) went one step further. He asked coders to develop independently a set of categories for a subsample of texts and to perform the coding. The results by the coders are compared[2]. Next the coders are given the opportunity to negotiate about the categories and the codings. This is to result in revised sets of categories. Now coding can be repeated and the process is repeated until the coders agree on the set of categories. In the mean time they have also established some coding rules.

Once the categories have been developed, coding can start. Not surprisingly, a coder coding according to the representational approach performs several tasks that can be compared to the tasks performed by a coder in qualitative research. With respect to the theme categories the following tasks are to be performed:

- The coder interprets a text block – in quantitative research a text block is at least on the level of a unit of analysis, in qualitative research the coder can take a text segment in a text block;
- The coder assigns the phenomena in this block to existing theme categories taking into account the context in which these phenomena occur – in the quantitative research the category is a factual code, in the qualitative research it might also be a referential code;
- If, according to the coder a phenomenon should be assigned to a category that does not exist yet, he or she adds this category (possibly after consulting the investigator) – in qualitative research the categories might also be changed or refined, etc.;
- If a search entry for a specific category does not exist, it is added – in qualitative research search entries are not relevant. In the open coding phase they might have served as codes.

This listing shows that coding in qualitative research and coding in quantitative research according to the representational approach have much in common. This is despite the fact that the goal in both types of research is different. At several places cooperation is possible. Fielding and Lee (1998) picture their project CAQDAS (Computer Assisted Qualitative Data Analysis Software) for investigating the software available for qualitative research. In the chapter on managing data they present lots of details on the way theme categories are developed and on the way coding is performed, both in an individual setting and as in teams.

Miles and Huberman (1994:64) emphasize the importance of pretesting and revising code books, because initial coding instructions often yield poor agreement. Carey et al (1996) testify that in their study two coders independently coded 320 text segments two times. The purpose of the first coding comparison was to pretest and remedy problems with the code book. Once these problems were identified and fixed, the same two coders used the amended code book to recode the same 320 passages of text. This let them estimate the final, or achieved, degree of intercoder agreement.

Reasons for the discrepancies in the codebook included problems such as redundant codes for the same belief, vague code definitions, lack of mutual

exclusivity between codes, or lack of shared understanding in the procedures for using specific codes.

Interjudge reliability

In case human coders perform the coding task, the investigator must realize that there are at least two coders and that their assignments are compared. This is to become informed about how similar they performed their tasks (How well is hard to say, as there is not by definition a correct answer.) Doing this comparison is not a decision moment for the investigator, it is a must.

The kind of reliability that is considered here is often denoted as agreement. There is a difference however. This is well described: "Interrater reliability provides an indication of the extent to which the variance in the ratings is attributable to differences among the rated objects. … Interrater agreement represents the extent to which the different judges tend to assign exactly the same rating to each object" (Tinsley & Weiss, 2000: 98).

Agreement is also considered as a special kind of association. Again there is a difference. It is important to determine the similarity of the content of behavior (in a broad sense) between coders in general with the degree of identity of this behavior. The behavior of one coder does not have to be predicted from that of the other. In the case of association one investigates the strength of the linear relationship between variables. Here the goal is to predict the values of one variable from those of the other. With regard to agreement, most important is the similarity of the content of behavior between coders, with the goal of determining the degree of identity of this behavior. The basic idea of an agreement index is looking at the fraction of units on which coders agree.

Consensus is growing among investigators who use indices for nominal scale agreement in their research that the preferred index should be of the type

$$I = \frac{O - E}{M - E},$$

where O stands for observed, E for expected, and M for maximum agreement.

The expected agreement is usually based on the marginal distributions. Coders in text analysis are usually equally skilled and this distribution is not known in advance. Therefore one takes the marginal distribution computed over all codings. It implies that only one stochastic variable is available. In this situation the term "agreement between codings" is used, instead of "agreement between coders." Hereafter only this situation of codings is considered. It implies that a situation, in which the marginal distribution per coder is used for computing the expected agreement, as is often found in medicine, is not dealt with. Actually in the present situation the individual coders are not relevant; they are supposed to come from a pool of coders which are all equally skilled and equally able to do the job.

A number of requirements to be posed to agreement indices have been formulated (Popping, 1988). The index that suffices best in the situation agreement in text analysis studies is to be computed, has been proposed by Scott (1955). This index was developed for the situation in which two coders have been used. Since it has been extended to all realistic research situations (Popping, 1992b). These research situations include:

- the number of judgments per unit (this number does not always have to be the same);

- use of weights based on the seriousness of disagreement between category pairs;
- the amount of agreement for a fixed category;
- the amount of (dis-) agreement for two different fixed categories.

In practice these situations do not appear separately, but in combination with each other.

The definition of agreement: pairwise (proportion of agreeing pairs of judgments), simultaneous (there is only agreement with respect to a unit, when all judgments are into the same category), or majority (there is only agreement with respect to a unit, when k out of the m judgments are into the same category) also has been subject of discussion. Popping (in press) has shown the pairwise agreement is to be used.

The general formulation for the agreement index as indicated before, based on Scott but covering all situations, is

$$K = (P_o - P_e) / (1 - P_e), \tag{1}$$

where P_o denotes the observed proportion of agreement, and P_e the proportion of agreement under the null hypothesis of independence.

Assume there are N units and c categories. Define n_{sj} as the number of times unit s has been assigned to category j. A matrix w is used for weights, w_{ij} is the weight for the situation in which one rating of an object is to category i and the other rating is to category j. In the standard situation one uses $w_{ii} = 1$ and $w_{ij} = 0$ (where $i <> j$). The algorithm $w_{ij} = 1 - |i-j|/(c-1)$ is used to indicate a linear weighted relation, and $w_{ij} = 1 - (i-j)^2/(c-1)^2$ refers to the squared weighted relation (Cicchetti, 1972). In practice these algorithms are used when the data are on an ordinal or interval level of measurement.

Now the number of times unit s has been assigned to category i can be computed:

$$n_{s.} = \sum_{i=1}^{c} n_{si.}. \tag{2}$$

This number must at least be 2, as otherwise no comparison to another coding can be made. The number allows computing the proportion of codings into a pair of categories:

$$p_{ii} = \sum_{s=1}^{N} n_{si} (n_{si} - 1) / \{N n_{s.} (n_{s.} - 1)\}, \tag{3}$$

for all i unequal j:

$$p_{ij} = \sum_{s=1}^{N} n_{si} n_{sj} / \{N n_{s.} (n_{s.} - 1)\}. \tag{4}$$

The total proportion of codings into category i is:

$$p_i = \sum_{s=1}^{N} n_{si} / N n_{s.}. \tag{5}$$

Now the observed and expected agreement over all categories can be defined:

$$P_o = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\, p_{ij} \; ; \tag{6}$$

$$P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}\, p_i\, p_j \; , \tag{7}$$

as well as the values for the single category i

$$P_{o(i)} = \sum_{j=1}^{c} w_{ij}\, p_{ij} / p_i \; ; \tag{8}$$

$$P_{e(i)} = \sum_{j=1}^{c} w_{ij}\, p_j \; . \tag{9}$$

This index covers nearly all empirical situations that are distinguished. The value the agreement index takes should meet some criterion. Generally this criterion is a rule of thumb as mentioned in literature, e.g. Landis and Koch (1977: 165) claim there is almost perfect agreement in case kappa ≥ .81. Here no attention is given to aspects of the actual investigation like for instance the difficulty of the coding task. This is to be regretted. The point is illustrated by looking at a study by Tilly. Tilly (1981: 74-75) contrasted the reliability of two questions. The first one concerned the occupation of respondents engaged in a strike, this is very concrete. The second question concerned the degree of coordination of the strike as a whole, this is very abstract. In the first situation 94% consistency was found between pairs of coders, in the second situation it was only 59% consistency.

The agreement statistic can be seen as a product variable. One also needs a process variable. This variable deals with the question how a specific degree of reliability is reached. This might be quite complex, the explanation is especially on coding rules. Popping and Roberts (in press) needed a complete article to explain how they did this in their study on modalities of democratic transformation in Hungary.

Coding can be a stepwise procedure, which has its effects on the reliability findings. In the study on democratic transformation we had to look at:

- Do coders agree on whether a text contains a modal clause?
- When a modal clause is identified, do coders identify the same one?
- Once coders have identified the same modal clause, do they agree on the clause's modality?
- Once coders have identified the same modal clause, do they agree on its rationale?

The following requirements are demanded from coders. They must

- be aware of the goal of the investigation;
- exactly know the meaning of the categories used;
- be trained in identifying the target behavior;
- be trained in which category to look first at (in case they have to code different categories at once);

- know which category is to be preferred in which situation, and why (in case there are reasons to assign different categories to a text unit).

The coder training is relevant, it is a two-phase activity. In terms of the above research is made clear as follows:

- First, each coder must separately identify modal clauses, rationales, and modality classifications in the same texts.
- Second, coders must jointly develop a set of coding rules for making these identifications. This second step requires considerable discussion (and consultation with their supervisors). Rules that emerge from these discussions must be explicit.

Here one must be careful. High coder agreement might be an effect of coder training, but, as Hak and Bernts (1996) showed, it is also possible that coders socialize each other in using practical rules of application that are not covered by the coding instruction. This should be prevented.

It is usually in the training phase that texts are coded by more than one coder, and that the final codings are compared. Some, e.g. Elder et al (1993: 60) assume that 10% of the data must be used in this training. I would like to have some systematic knowledge regarding this issue. We know very little about coder training.

Validity deals with the question whether the findings in the investigation represent real events. The investigator is especially confronted with the issue when the categories and search entries are developed and applied. In the development phase the main issue is first of all whether the categories that are defined are the relevant ones for the research question. Next comes the question whether the correct method of text analysis is applied. In case the thematic approach will be followed, the investigator has to wonder whether the correct search entries are used.

The questions about the categories and the search entries can also be posed in the coding phase, especially when the representational approach is followed. To be added is the question whether the search entries are used in a correct way.

Conclusion

An overview is given of how text analysis can be applied to the answers on open-ended questions in a public opinion poll or survey. A number of decisions to be taken by the investigator are discussed and the need of computing intercoder reliability is emphasized.

When the coding task has been performed and the data-matrix based on this coding is available, this data-matrix must be merged with the data-matrix containing the data from the survey. The identification number serves to make that the correct data are combined. The number of rows in the data matrix based on the text analysis must be equal to the number of respondents in the study. This implies that in case the text analysis has been on the level of recording units, some units have to be taken together in order to arrive at the correct number of observations. How this is done depends on the type of data that is to be aggregated now. Sometimes one might have to use the sum of the scores, at another time the mean of the highest score.

Notes

1. It is possible to overcome this problem by using knowledge from artificial intelligence. In the earlier mentioned studies on classifying occupations (Popping, 1995, 1997), trigrammes were used. The word under investigation is divided in

parts of for example 3 characters and the proportion of parts it counted that match with parts found in the same way from words in the dictionary. The word "baker" is divided into "__b", "_ba", "bak", "ake", "ker", "er_", "r__" (the underscore "_" should be read as a space).

2. Intercoder agreement might be computed in the situation in which coders each developed a set of categories and also performed the coding (Popping, 1983).

References

Carey, J.W., Morgan, M. & Oxtoby, M.J. (1996). "Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research." *Cultural Anthropology Methods*, 8 (3): 1-5.

Cicchetti, D.V. (1972). "A new measure of agreement between rank ordered variables." *American Statistical Association*, Proceedings of the 80th annual convention, 7, 17-18.

Contas, M.A. (1992). "Qualitative analysis as a public event: The documentation of category development procedures." *American Educational Research Journal*, 29 (2): 253–66.

Elder, G.H. Jr., Pavalko, E.K. & Clipp, E.C. (1993). *Working with Archival Data. Studying Lives*. Newbury Park: Sage.

Geer, J.G. (1988). "What Do Open-Ended Questions Measure?" *The Public Opinion Quarterly*, 52 (3): 365-371 .

Hak, T. & Bernts, T. (1996). "Coder training: Theoretical training or practical socialization?" *Qualitative Sociology*, 19 (2): 235-57.

Heinrich, H.A. (1996). "Traditional versus computer aided content analysis. A comparison between codings done by raters as well as by Intext." In: Faulbaum, F. & Bandilla, W. (eds.), *SoftStat'95. Advances in Statistical Software 5*. Stuttgart: Lucius & Lucius: 327–333.

Holsti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities*. London: Addison Wesley.

Kurasaki, K.S. (2000). "Intercoder reliability for validating conclusions drawn from open-ended interview data." *Field Methods*, 12 (3): 179-194.

Landis, J.R. & Koch, G.G. (1977). "The measurement of observer agreement for categorical data." *Biometrics*, 33 (1): 159-174.

Lofland, J. (1971). *Analyzing Social Settings*. Belmont, CA: Wadsworth Publishing Company.

McDonald, C. (1982). "Coding open–ended answers with the help of a computer." *Journal of the Market Research Society*, 24 (1): 9–27.

Miles, M.B. & Huberman, A.M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: Sage.

Montgomery, A.C. & Crittenden, K.S. (1977). "Improving coding reliability for open-ended questions." *Public Opinion Quarterly*, 41 (2): 235-243.

Popping, R. (1983). "Traces of agreement. On the Dot-product as a coefficient of agreement." *Quality and Quantity*, 17 (1): 1-18.

Popping, R. (1988). "On agreement indices for nominal data." In: Saris, W.E., & Gallhofer, I.N. (eds.), *Sociometric research: Data collection and scaling*. Hampshire: McMillan: 90-105.

Popping, R. (1992a). "In search for one set of categories." *Quality & Quantity*, 25 (1): 147–55.

Popping, R. (1992b). *Taxonomy on Nominal Scale Agreement 1945–1990*. Groningen: iec *Pro*GAMMA.
Popping, R. (1995). "The performance of a knowledge-based system for converting occupational classifications." *Quality and Quantity*, 29 (3): 273-286.
Popping, R. (1997). "Reliability of registrations: A feasibility study into registration of occupational and educational titles in hospitals." *Quality and Quantity*, 31 (3): 305-315.
Popping, R. (2000). *Computer-assisted Text Analysis*. London: Sage.
Popping, R. (in press). "Some views on agreement to be used in content analysis studies." *Quality & Quantity*, in press.
Popping, R. & Roberts, C.W. (in press). "Coding issues in semantic text analysis." *Field Methods*, in press.
Riffe, D., Lacy, S. & Fico, F.G. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis in Research.* Mahwah, NJ: Lawrence Erlbaum.
Roberts, C.W. & Popping, R. (1993). "Computer supported content analysis. Some recent developments." *Social Science Computer Review*, 11 (3): 283-291.
Scott, W.A. (1955). "Reliability of content analysis: The case of nominal scale coding." *Public Opinion Quarterly*, 19 (3): 321-5.
Shapiro, G. (1997). "The future of coders: Human judgments in a world of sophisticated software." In: Roberts, C.W. (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates: 225–238.
Tilly, C. (1981) *As Sociology meets History*. New York: Academic Press.
Tinsley, H.E.A. & Weiss, D.J. (2000). "Interrater reliability and agreement." In: Tinsley, H.E.A. & Brown, S.D. (eds.), *Handbook of applied multivariate statistics and mathematical modeling*. San Diego (etc): Academic Press: 95-124.
Van der Eijk, C., Irwin, G.A. & Niemöller, B. (1988). *Dutch parliamentary election panel-study 1981-1986*. Amsterdam: Steinmetz Archives, Dutch Interuniversity Election Workgroup.