4
Different Functions of Coding in the Analysis of Textual Data

John Seidel and Udo Kelle

Now it was early afternoon. Gorman had met him as requested, at the Mexican Water trading post. They'd made the bone-jarring drive back into Chichinbito Canyon country. Rather quickly Officer Gorman had proved he was the sort of man who— as Leaphorn's grandmother would have said— counted the grass and didn't see the grazing.

Gorman was sitting now in Leaphorn's car, waiting (uneasily, Leaphorn hoped) for Leaphorn to finish whatever the hell Leaphorn was doing. What Leaphorn was doing was looking past the grass at the grazing... (from Tony Hillerman's detective novel Skinwalkers)

Two functions of codes

Coding is of crucial importance for computer-aided qualitative data analysis. Although codes represent the decisive link between the original 'raw data', that is, the textual material such as interview transcripts or field notes on the one hand, and the researcher's theoretical concepts on the other, most of the methodological literature does not concentrate much on coding but rather on how to work with codes and coded material.

In this chapter we will discuss two different modes of coding1 in the analysis of textual data:

1. A code can denote a text passage containing specific information in order to allow its retrieval. This mode of coding is analogous to the construction of a book index: the word 'Popper' followed by a page number informs the reader that the philosopher Karl Popper is mentioned on this page. When coding an open interview the code political affiliation together with certain line numbers informs the analyst that in the text passage with these line numbers the respondent talks about political affiliation.

2. A code can also serve to denote a fact. Thus, a text segment where an interviewee expresses his predilection for the Liberal Party can be denoted by the code Liberal Party affiliation and the analyst can then draw on the information that the respondent is an adherent of the Liberal Party without even looking at the original text.

One can call these two forms of coding indexing and summarizing, or one can, by drawing on a distinction made by Richards and Richards (in this volume, p. 83), differentiate between the factual and the referential functions of code categories. But one must be careful not to reify this distinction: codes themselves are neither factual nor referential. Furthermore, coding always has both functions: although one can drop the reference of a code to a certain text segment, a code can never lose its referential function. There must always be some reference to some unit of analysis (an interview, a respondent or something else), otherwise any meaningful information will be lost (for example, about which member of the sample is a Liberal). On the other hand a mere index also contains factual information (for example, that the author of the book mentions certain people) even if this information is not very rich in itself. Consequently it is not the specific type of codes — 'factual' or 'referential' — but rather their use in a certain setting that determines which of the two functions plays the dominant role.

Codes as representations of phenomena

Classical content analysis (see Berelson, 1952; Krippendorf, 1980), a method for the analysis of unstructured data that has a long tradition in mass communication research, can in some respects be regarded as paradigmatic for a coding strategy whereby codes serve as representations of the investigated phenomena. A researcher using this method who is, for example, interested in comparing the commentaries on certain political events by different newspapers will proceed by condensing the information contained in the raw data to a minimum. For this purpose the unit of analysis (for example, sentences or paragraphs) has to be determined and a precise coding scheme has to be constructed. Then, every unit of analysis can be investigated as to:

1. whether it refers to one of the defined political events, and if so
2. which attitudes towards this event the author expresses, for example affirmation, disapproval or something in between.

By applying this method a code represents a value of a certain variable, for example the value 'affirmation' of the variable 'Evaluation of political events'. The code is then attached to a unit of analysis, such as a single sentence. Each appearance of a certain code represents a certain event that is of interest to the researcher, for example an affirmative attitude expressed by a newspaper reporter towards a specific political event. In classical content analysis, coding is usually followed by information reduction: the information provided by coding is used for the construction of a new (quantitative) data corpus that can be analysed with statistical procedures, for example a matrix which...
contains information about how different newspapers (the rows of the matrix) comment on certain political events (the columns of the matrix). Thus, for this kind of analysis it is crucial to focus almost exclusively on the codes and not on the raw data. But this is only possible if the codes can be seen as true representations of certain facts described by the raw data. Consequently the following requirements have to be fulfilled:

1. There has to be a high degree of certainty that the codes have been applied in a systematic and consistent way; in other words, the coding must have a high degree of validity and reliability: it is essential that every text passage that is coded as affirmation of political event A in fact contains a positive utterance about this event.

2. The coding of the raw data must be inclusive and exhaustive; this means one must be certain that every single instance of the investigated phenomenon that occurs in the raw data is coded. This is of special importance if, by the employment of statistical procedures, the absolute or relative frequency of the occurrence of codes are used to draw inferences, for example for the support or rejection of hypotheses.

These requirements make it essential that whenever the kind of analysis is employed in which codes are used as a condensed representation of the facts described by the data, a precise coding scheme is developed before coding starts, since:

1. For pragmatic reasons alone, inclusive and exhaustive coding would not be possible if the researcher did not have a ready-made category scheme to hand right from the start. If, instead, the coding scheme was being permanently altered, it would be necessary to permanently re-code the previously coded data with the newly developed categories.

2. Objective, and therefore reliable, coding can only be conducted if all coders employ exactly the same coding scheme.

Consequently, an analysis strategy that relies mainly on the factual function of codes, that is where codes are used as representations of phenomena, requires a deductive approach: the relevant variables and their values (that form the codes) have to be determined before data are coded.

Codes as heuristic devices

However, this requirement for a deductive approach would cause severe difficulties in the context of a qualitative research strategy. Let us take a short look at the theoretical roots of qualitative research to clarify this point. Although there is a puzzling heterogeneity of qualitative approaches (see Tesch, 1990: 55ff.), it has often been emphasized that most of these approaches are at least implicitly based on a common underlying concept of human action, which has been referred to using notions like The Interpretive Paradigm (Wilson, 1970), Interpretive Sociology (Giddens, 1976) or Interpretive Interactionism (Denzin, 1989).

According to the interpretive paradigm the meaning of human action and interaction can only be adequately understood if the interpretations and the common-sense knowledge of the actors are taken into account. This theoretical postulate has far-reaching methodological consequences: the researcher must be able to gain access to the interpretations and the common-sense knowledge of the members of the social world investigated. If his or her goal is to describe their actions adequately, which is with respect to the meaning these actions have for the actors, he or she must be able to perceive the world to a certain extent in the same way as they do. Or, as Thomas and Znaniecki put it in their early study of The Polish Peasant: ‘We must put ourselves in the position of the subject who tries to find his way in this world, and we must remember, first of all, that the environment by which he is influenced and to which he adapts is his world and not the objective world of science’ (1958: 1846f.).

This demand for ‘empathic understanding’ of or access to the common-sense knowledge of the investigated form of social life makes it difficult, if not impossible, to employ a hypothetico-deductive (H-D) research strategy, since this would require the development of useful hypotheses before collecting empirical data. Instead, if one wants to learn something about the actor’s point of view one has first to enter the empirical field. Meaningful hypotheses can be established only after gathering data, that is after establishing contact with the people in this field through interviewing or observation.

Hence, the goal of discerning the specific perspectives, world-views, 'local knowledge' (Geertz, 1973), etc. of the members of a specific yet unknown form of social world, can hardly be achieved by means of an H-D strategy: instead of constructing hypotheses before entering the field, the researcher has to go out and look at how people perceive and interpret their world themselves.

For that reason qualitative enquiry in most cases starts with observation, recording, listening, etc., which means by collecting sometimes large amounts of unstructured data, and then hypotheses and theories are developed on the basis of this material. If one wants to discuss the role of coding in this research paradigm it is helpful to distinguish between three operations in this process: (1) noticing relevant phenomena, (2) collecting instances of these phenomena, (3) analysing
these phenomena in order to find commonalities, differences, patterns, structures, etc.

**Noticing relevant phenomena**

Noticing relevant phenomena is the central task of the researcher who enters an empirical field without ready-made hypotheses. Instead of searching for supporting or disconfirming evidence for a hypothesis, one tries to discover things that one has not previously expected. Of course this must not seduce us into thinking of the researcher's mind as a tabula rasa. An open mind does not mean an empty head; the researcher always brings some theoretical preconceptions about empirical facts. Rather, they should be referred to as (partly implicit) conceptual networks that provide us with some 'lenses' for the perception of the empirical world. We would like to address these conceptual networks as perspectives rather than as 'hypotheses' or 'theories'. These perspectives help the researcher to select relevant phenomena, and of course researchers with different perspectives will select different phenomena. But behind these differences is the simple principle in qualitative research of going out into the empirical world and noticing relevant phenomena. In the first step this translates into making observations, writing field notes, tape-recording interviews, etc., by which a general record of what one has noticed is produced.

The next step of noticing would be to identify the relevant phenomena which are contained in these general records. Identifying phenomena within one's field notes, protocols or interviews quite often (but not always!) takes the form of 'coding' (see the Introduction), which is in fact a procedure with a long tradition in the 'folklore' of qualitative research. One of the first descriptions of this process can be found in an article by Howard Becker and Blanche Geer:

A systematic assessment of all data is necessary before we can present the content of a perspective... We have tentatively identified, through sequential analysis during the field work, the major perspectives we want to present and the areas... to which these perspectives apply. We now go through the summarized incidents, marking each incident with a number or numbers that stand for the various areas to which it appears to be relevant. This is essentially a coding operation... its object is to make sure that all relevant data can be brought to bear on a point. (Becker and Geer, 1960: 280-281)

It should be noted that Becker and Geer outline a quite different process than that described earlier in this chapter. The incidents which are coded do not represent instances or examples of a general phenomenon or fact named by a code; the code only refers in a quite vague manner to one of 'the various areas to which it appears to be relevant'. The purpose is not to condense the information which is relevant for the researcher with the objective of creating a quantitative data matrix, but to 'make sure that all relevant data can be brought to bear on a point'. Here the function of codes is clearly restricted to 'signposting' - they are stored together with the 'address' of a certain text passage. Drawing on this information the researcher can then locate all the possible information provided by the data on the relevant topic. Thus coding is a necessary preparation for the two other operations that have already been mentioned: collecting instances of relevant phenomena and comparing relevant phenomena.

**Collecting instances of relevant phenomena**

After coding, the material can be broken into parts and then collected according to the topics which served as codes. Many texts on ethnographic or qualitative methods address this process of 'reassembling' and 'disassembling' the data, in which another widely applied folklore technique - known as 'cut-and-paste' - is often used: the text material is cut into small pieces according to the topic-oriented codes, glued onto index cards and then sorted or put into a file folder (see, for example, Freidson, 1975: 271; Lofland and Lofland, 1984: 134; Taylor and Bogdan, 1984: 136). In doing so the text segments are taken 'out of the chronological narrative form of your field notes or interview write-ups and into a storage system where you can easily order and retrieve it' (Lofland and Lofland, 1984: 132). However, this mindless and mechanical procedure that helps to collect relevant phenomena is nothing but a preliminary for the central procedure of qualitative text interpretation.

**Analysing relevant phenomena**

This crucial procedure is making sense of the data by analysing the relevant phenomena found in the material and collected by the mechanical operation of cutting and pasting. This is conducted by comparing the different pieces of data in order to find commonalities, differences or linkages between them. The purpose is 'the construction of meaningful patterns of facts' (Jorgenson, 1989: 107), by looking for structures in the data. To some degree this process is similar to doing a jigsaw puzzle: as with a jigsaw puzzle the analyst would start by collecting certain pieces which are similar in a certain respect. If the picture is of a landscape comprising a tree, a building and the sky, one could at first use 'Tree', 'Building' and 'Sky' as 'codes' and sort pieces according to these categories. At this point, more refined inspections of the pieces will be necessary. Shapes and colours will be meticulously examined and compared. One looks for features of the pieces that give hints as to the linkages between them. This is a similar process to that in qualitative data analysis: once the analyst has sorted the pieces that
were allocated to a certain category he or she has to study the text segments as things in themselves. He or she will analyse several parts of them and their connections, that is the specific way they could be linked or connected to form a meaningful picture. But unlike doing a jigsaw puzzle, the researcher analysing qualitative data cannot refer to the picture on the box. He or she has no model or blueprint at hand to identify the features which would help him or her to establish linkages between the pieces. Using a model or blueprint would be how a researcher in the H-D model proceeds: he or she would construct a theoretical model that provides a template, to indicate which information is relevant, before analysing data. In contrast, the qualitative researcher employing an interpretive methodology tries to derive this information from his or her data. In many qualitative approaches, especially in those with strong roots in hermeneutical philosophy or phenomenology, for example discourse analysis, this would be conducted by means of a thorough fine-grained analysis of certain text segments. Thereby, hitherto unknown aspects of the phenomenon under study are discovered through careful and intensive inspection of the ‘raw data’, that is, the original text. These are the aspects that help to combine the pieces of information in a new way in order to gain a new image of the investigated empirical domain.

The codes employed in such an analysis strategy are imprecise and vague, if compared with codes used for the representation of facts. They will be attached not to precisely defined incidents in the data but to text segments ‘which seem [...] to be distinct incidents, anecdotes, or stated opinions about discrete events’ and which are ‘tentatively [...] classified into the simple [...] content categories we had decided in advance’ (Freidson, 1975: 271). But this lack of precision is a necessary prerequisite for their use in interpretive research, which incorporates a methodology of discovery. Here, in contrast to a methodology of hypothesis testing, the researcher must not restrict the scope of the investigation in advance by determining precise categories, since the goal is not to recover certain already known phenomena in the empirical field but to discover new ones. Coding is nothing more than a heuristic device for discovery.

Confusing the different functions of codes and the resulting methodological problems

Computer-assisted methods of qualitative data analysis were initially developed to assist researchers who used codes as heuristic devices. The first software packages represented mere electronic ‘cut-and-paste’ tools, but at present more and more powerful facilities for the manipulation of codes are being added to many of the programs:

1. tools that enhance the construction of ‘networks’ or ‘hierarchical trees’ of code categories (see Part II),
2. knowledge-based systems that help to retrieve information about how text segments attached to certain codes are distributed throughout the whole document, whether they overlap or appear at certain distances (Part III),
3. devices for the quantitative analysis of codes and of their distribution across documents (Part IV).

These facilities offer fascinating new possibilities for analysts to ‘play’ with their data and thereby help to open up new perspectives and to stimulate new insights. They can also help to combine qualitative with quantitative methods or an H-D approach with interpretive research strategies. But these possibilities also contain specific dangers, because the same technical tool can be used for two totally different research strategies which employ codes differently: as indexes for text segments that are coded and retrieved by an electronic ‘cut-and-paste’ device on the one hand, and as representations of factual information contained in the raw data on the other. Therefore analysts can—without realizing—confuse the two modes of coding: they can involuntarily switch from using the referential function of codes (that means from collecting text segments that refer in a broad and general way to a number of somewhat vaguely defined concepts) to treating codes as if they were representations of factual information.

We will call this the danger of losing the phenomenon by reifying the code: the analyst starts to work exclusively on his or her codes and forgets about the raw data, although the necessary prerequisite for doing so has not yet been fulfilled: there is only a loose coupling between a code and a piece of data instead of a well-defined relation between a code and a phenomenon, since the code was not attached to denote a certain discrete event, incident or fact, but only to inform the analyst that there is interesting information contained in a certain text segment, related to a topic represented by a code.

But alongside the danger of losing the information contained in the text segment, there is also a second danger that arises when the two modes of coding are confused: losing the context of the information.
Like Officer Gorman in Hillerman’s detective story the analyst will then start ‘to count the grass and not see the grazing’.

What precautionary measures can be taken to avoid these two dangers of losing the phenomenon and losing its context? Generally speaking there are two strategies: one corresponds to the H-D model, the other is related to an interpretive research strategy.

1. If the H-D model is applied one would have to ensure that every code is suitable for representing a distinct and objectively verifiable fact whose presence could be ascertained independently from a specific context. Additionally, the analyst would have to ensure that the codes are applied consistently and in a reliable manner to the raw data. In particular one has to pay attention to ensuring that the coding is inclusive and exhaustive, that is, that every incident which is represented by a certain code category is in fact coded. Since there is already much information available about the reliability and consistency of category schemes in the technical literature on content analysis, we do not need to go into further detail. But it should be mentioned that if a researcher starts with an exploratory strategy of text analysis and then switches to an H-D mode, that means that in practice he or she starts with referential coding and then switches to factual coding, it would, in almost all cases, be necessary to (1) reformulate the coding scheme and (2) re-code material.

2. In interpretive research there are quite different strategies to cope with the problems of losing the phenomena and losing the context. Since the onset of an interpretive research tradition there has been a distinctive awareness among its adherents of the context-relatedness of the investigated social phenomena, which has inspired much criticism of the H-D model of research (see Cicourel, 1964). Consequently, there has always been concern among qualitative researchers that by applying cut-and-paste techniques ‘the totality of philosophy as expressed by the interviewee – which is closely related to the goal of the study’ (Wiseman, 1979: 278) can be destroyed. Usually two strategies are recommended to avoid losing the context of the text segments: (1) analysts are encouraged to acquaint themselves, by careful reading, with the entire case or document as, for example, Agar suggests: ‘read the transcripts in their entirety several times. Immerse yourself in the details, trying to get a sense of the interview as a whole’ before breaking it into parts (Agar, 1980: 103). (2) Several methods are proposed that help to restore or find the original context at any time: it is recommended that interviews are typed in duplicate and that one copy is always left ‘intact to be read in its entirety’ (Wiseman, 1979: 278). Other practitioners advise the researcher to ‘include enough of the context to understand the data fully’ (Taylor and Bogdan, 1984: 137). Another proposal is to include on each data fragment the ‘identification in code of the persons being interviewed, the interview number, and the page number of the typescript so that the context of the selection could be returned to and examined’ (Freidson, 1975: 270).³

These strategies of keeping in contact with the raw data can not only be applied to avoid the danger of losing the context, but also to diminish the menace of reifying the codes and losing the phenomenon.

Both methodological approaches – trying to guard the reliability, consistency and inclusiveness of code use, and trying to keep contact with the raw data – address the problem of whether the researcher’s theoretical concepts (which should not be confused with codes as technical devices) are really grounded in the raw data, or in other words, the problem of validity. But they do so in quite different ways in the context of quite different models of the research process.

The purpose of this chapter is not to argue in favour of one model as the only one that is appropriate. Nor do we want to revive old debates from the 1970s between true believers in a certain methodological credo and heretics, or to claim that there are insurmountable barriers between exploratory and H-D modes of text analysis. On the contrary, the combination of research strategies can have quite a fruitful impact on the research process if, for example, systematic exploratory procedures for empirical hypothesis generation are employed and these hypotheses are then tested by means of hypothetico-deductive methods. Nonetheless, this combination can have harmful results if it is not conducted on the basis of careful methodological considerations, in which the necessary prerequisites for the application of a certain method are taken into account. Furthermore, it must be ensured that the chosen mode of coding or textual analysis fits into the framework of the research setting and the research question.

Notes

1. I am particularly grateful to Tom Richards and Lyn Richards, whose distinction between referential categories and factual categories inspired our thinking on the different modes of coding. [Udo Kelle]

2. Thanks to Ian Dey for this nice aperçu.

3. Of course it is much easier to retain the original context when using computer-aided methods of coding and retrieval than with cut-and-paste techniques, since the text segments are usually not actually ‘cut out’, but pointers are merely attached to them. But, most interestingly, there are only a few programs that provide an easy and user-friendly facility for retrieving text segments in context.