8.3.3. Coding Qualitative Data

Just as with quantitative research, qualitative researchers code data by identifying patterns. However, in qualitative research, coding is usually grounded in the data. In other words, the schemes for qualitative coding generally emerge from the data rather than being decided on and preimposed prior to the data being collected or coded. This process, in which initial categories are based on a first pass through the data, is sometimes known as open coding. Qualitative researchers explore the shape and scope of the emerging categories and investigate potential connections among categories. As more data are coded, researchers also consider aspects such as the range of variation within individual categories. These processes can assist in the procedure of adapting and finalizing the coding system, with the goal of closely reflecting and representing the data.

For example, one way of coding qualitative data can involve examining the data for emergent patterns and themes, by looking for anything pertinent to the research question or problem, also bearing in mind that new insights and observations that are not derived from the research question or literature review may be important. Paraphrases, questions, headings, labels, or overviews can be assigned to chunks of the data. These labels or indicators are usually not precise at the early stages. The data, rather than the theory or framework, should drive the coding. Many researchers try to code the data by reminding themselves that they will need to explain how they arrived at their coding system, keeping track of the data-based origins of each of their insights. Interesting data that are extra to the goals of the study are not discarded; they are kept in mind and possibly also coded. Themes and topics should emerge from the first round of insights into the data, when the researcher begins to consider what chunks of data fit together, and which, if any, are independent categories. Finally, a conceptual schema or organizational system should emerge, by which researchers consider their contribution to the field. At this stage, researchers often ask themselves if they can tell an interesting narrative based on the themes in the data. At this stage they are often ready to talk through their data and the patterns with others, so that input can help them in the stages before they write up their research. This is just one method by which qualitative researchers can code and analyze their data. Denzin and Lincoln (1994) presented a comprehensive picture of the many alternatives.

One problem with developing highly specific coding schemes is that it can be problematic to compare qualitative coding and results across studies and contexts. However, as Watson-Gegeo (1988) pointed out, although it may not be possible to compare coding between settings on a surface level, it may still be possible to do so on an abstract level. Whereas a particular event may not occur in two settings, the same communicative need can exist in both. For example, in examining the relationship between second language learning and attitudes of immigrant children, although one study may focus on the school context and another on the home context, and each may examine different types of events in the data, the overall questions and answers may be comparable.

8.4. INTERRATER RELIABILITY

Regardless of the choice researchers make from the wide range of different types of data coding that are possible, establishing coding reliability is a crucial part of the process. The choice of which coding system to adopt, adapt, or devise ultimately depends on the researcher's goals and the type of study being carried out. However, it is common to ensure that the coding scheme can be used consistently or reliably across multiple coders wherever possible. This is known as interrater reliability, a concept introduced in chapter 4.

Because coding involves making decisions about how to classify or categorize particular pieces of data, if a study employs only one coder and no intracoder reliability measures are reported, the reader's confidence in the conclusions of the study may be undermined. To increase confidence, it is important not only to have more than one rater code the data wherever possible, but also to carefully select and train the raters. It may be desirable to keep coders selectively blind about what part of the data (e.g., pretest or posttest) or for which group (experimental or control) they are coding, in order to reduce the possibility of inadvertent coder biases. In some cases, researchers act as their own raters; however, if, for example, a study involves using a rating scale to evaluate essays from second language writers, researchers may decide to conduct training sessions for other raters in which they explain something about the goals of the study and how to use the scale, provide sample coded essays, and provide opportunities and sample data for the raters to practice rating before they judge the actual data. Another way to increase rater reliability is to schedule coding in rounds or trials to reduce boredom or drift, as recommended by Norris and Ortega (2003). One question that is often raised is how much data should be coded by second or third raters. The usual answer is, as much as is feasible given

the time and resources available for the study. If 100% of the data can be coded by two or more people, the confidence of readers in the reliability of the coding categories will be enhanced, assuming the reliability scores are high. However, researchers should also consider the nature of the coding scheme in determining how much data should be coded by a second rater. With highly objective, low-inference coding schemes, it is possible to establish confidence in rater reliability with as little as 10% of the data. We now turn to a discussion of those scores.

8.4.1. Calculating Interrater Reliability

In addition to training the raters and having as much data as possible scored by more than one rater, it is also crucial to report interrater reliability statistics and to explain the process and reliability estimate used to obtain these statistics.

8.4.1.1. Simple Percentage Agreement

Although there are many ways of calculating interrater reliability, one of the easiest ways is through a simple percentage. This is the ratio of all coding agreements over the total number of coding decisions made by the coders. For example, in Mackey and Oliver's (2002) study of children's ESL development, both researchers and one research assistant coded all of the data. This process yielded an interrater reliability percentage of 98.89%, meaning that there was disagreement over only 1.11% of the data. Simple percentages such as these are easy to calculate and are appropriate for continuous data (i.e., data for which the units can theoretically have any value in their possible range, limited in precision only by our ability to measure them—as opposed to discrete data, whose units might, for example, be limited to integer values). Their drawback is that they have a tendency to ignore the possibility that some of the agreement may have occurred by chance. To correct for this, another calculation is commonly employed—Cohen's kappa (Cohen, 1960).

8.4.1.2. Cohen's Kappa

This statistic represents the average rate of agreement for an entire set of scores, accounting for the frequency of both agreements and disagreements by category. In a dichotomous coding scheme (e.g., coding forms as targetlike or nontargetlike), Cohen's kappa requires that the researcher determine how many forms both raters coded as targetlike, how many were

coded as targetlike by the first rater and as nontargetlike by the second, how many were coded as nontargetlike by the first and as targetlike by the second, and so on. The final calculation of kappa therefore involves more detail on agreement and disagreement than simple percentage systems, and it also accounts for chance.

8.4.1.3. Additional Measures of Reliability

Other measures, such as Pearson's Product Moment or Spearman Rank Correlation Coefficients, may also be used to calculate interrater reliability. These latter two are based on measures of correlation and reflect the degree of association between the ratings provided by two raters. They are further discussed in chapter 9, in which we focus on analysis.

8.4.1.4. Good Practice Guidelines for Interrater Reliability

In most scientific fields, including second language research and associated fields such as education, "there is no well-developed framework for choosing appropriate reliability measures" (Rust & Cooil, 1994, p. 2). Although a detailed examination and comparison of the many different types of interrater reliability measures is beyond the scope of this chapter (for more comprehensive reviews see Carmines & Zeller, 1979; Chaudron, Crookes, & Long, 1988; Gwet, 2001; Pedhazur & Schmelkin, 1991), general good practice guidelines suggest that regardless of which measurement is chosen, researchers should state which measure was used to calculate interrater reliability, what the score was, and, if there is space in the report, briefly explain why that particular measure was chosen. Some researchers also explain how data about which disagreements arose were dealt with; for example, if agreement was eventually reached and the data were included in the analysis, or if data (and how much) were discarded.

There are also no clear guidelines in the field of second language research as to what constitutes an acceptable level of interrater reliability. The choices and decisions clearly have lower stakes than, for example, in the field of medicine. However, the following rough guidelines based on rigorous standards in some of the clinical science research may be of some assistance (Portney & Watkins, 1993):

- For simple percentages, anything above 75% may be considered "good," although percentages over 90% are ideal.

- For Cohen's kappa, 0.81 to 1.00 is considered "excellent." In general, a reader should be concerned with a percentage of less than 80%, because this may indicate that the coding instrument needs revision.

8.4.1.5. How Data Are Selected for Interrater Reliability Tests

As noted earlier, in some second language studies the researchers code all of the data and calculate reliability across 100% of the dataset. However, an alternative is to have the second or third rater code only a portion of the data. For instance, in some studies the researcher may semi-randomly select a portion of the data (say 25%) and have it coded by a second rater (and sometimes by a third or fourth rater as well, depending on the size of the dataset and the resources of the researcher). If this approach is taken, it is usually advisable to create comprehensive datasets for random selection of the 25% from different parts of the main dataset. For example, if a pretest and three posttests are used, data from each of them should be included in the 25%. Likewise, if carrying out an interrater reliability check in an L2 writing study, essays from a range of participants at a range of times in the study should be selected.

It is often necessary to check intrarater reliability as opposed to interrater reliability. In Philp's (2003) study, she coded all of the data. She then recoded 15% of the data herself 6 months later to check for intrarater reliability. Intrarater reliability refers to whether a rater will assign the same score after a set time period. Philp used this system together with a standard check for interrater reliability, also having one third of her treatment transcripts double-coded by six assistants.

8.4.1.6. When to Carry Out Coding Reliability Checks

It is important to realize that if a researcher codes 100% of a dataset him- or herself, and then realizes that the coding system is unreliable, a great deal of unnecessary effort will have been expended, because the coding system may need to be revised and the data recoded. For this reason, many researchers decide to use a sample dataset (perhaps a subset of the data, or data from the pilot test) to train themselves and their other coders, and test out their coding scheme early on in the coding process. Following this initial coding and training, coders may then code the rest of the dataset independently, calculating interrater reliability at the end of the coding process on the data used for the research, rather than for the training.

When space permits, we recommend the following reporting on coding:

- What measure was used.
- The amount of data coded.
- Number of raters employed.
- Rationale for choosing the measurement used.
- Interrater reliability statistics.
- What happened to data about which there was disagreement (e.g., recoded? Not included?).

Complete reporting will help the researcher provide a solid foundation for the claims made in the study, and will also facilitate the process of replicating studies. If a low interrater reliability statistic is reported, this may be an indication that future studies will need to revise the coding system.

8.5. THE MECHANICS OF CODING

After selecting or devising an appropriate coding system, the researcher must determine how to go about coding the data. Implementations of systems vary among researchers according to personal preferences. Some researchers, for example, may prefer a system of using highlighting pens, working directly on transcripts, and marking such things as syntactic errors in one color pen and lexical errors in another, with a tally on each page and a final tally on the first page of the transcript. Other researchers, depending on their particular questions, may decide to listen to tapes or watch videotapes without transcribing everything; they may simply mark coding sheets when the phenomena they are interested in occur, and may decide to transcribe only interesting examples for their discussions. This system may also be used for written data, for which coding sheets are marked directly without marking up the original data. Still other researchers may prefer to use computer programs to code data if their research questions allow it. For example, if a researcher is interested in counting the number of words in different sections of an essay or selecting the central portion of a transcript for analysis, it would be much easier to use a word processor than it would be to do this exercise by hand. If the research questions relate to computer-assisted language learning, many CALL programs automatically record each keystroke a learner makes, and these data can easily be sorted and coded. Likewise, if the researcher wishes to focus on such things as reaction times to certain presentations on a computer screen, or eye movements as learners read and write text, reaction time software would be a possible choice.

8.5.1. How Much to Code?

As suggested previously, not all research questions and coding systems require that an entire dataset be coded. When selecting and discussing the data to code, researchers first need to consider and justify why they are not coding all their data. A second important step is determining how much of the data to code. This process is sometimes known as data sampling or data segmentation. Some researchers may decide it is important to code all of the data, whereas others may decide their questions can be answered by examining a portion of the data. In making decisions about how much and which portions of data to code, another point to consider is that the data to be analyzed must always be representative of the dataset as a whole and should also be appropriate for comparisons if these are being made. For example, if a researcher chooses to code the first 2 minutes of oral data from a communicative task carried out by one group of students and the last 2 minutes from another group of students, the data might not be comparable because the learners could be engaged in different sorts of speech even though the overall task is the same. In the first 2 minutes they might be identifying and negotiating the problem or activity, whereas in the final 2 minutes they might be making choices about how to complete the activity, or even how to communicate the outcomes to others. Another possibility is that the learners may begin to lose interest or feel fatigued by the end. In view of these concerns, the researcher could instead choose to take the middle section of the data—for example, the central 50 exchanges. Whenever possible, if only a portion of the data is being coded, researchers should check that the portion of data is representative of the dataset as whole. Of course, as with everything else about research design, the research questions should ultimately drive the decisions made, and researchers need to specify principled reasons for selecting data to code.

In much of Oliver's work (1998, 2000, 2002), she has made a general practice of coding only the first 100 utterances of each of her extended interactions. For example, in Oliver (2000) she coded the first 100 utterances of each teacher-fronted lesson and of each of the pair-work tasks in her study. Oliver made this decision because, in some cases, the interactions were only a little more than 100 utterances, and she needed a minimum comparable number of units to code. In summary, depending on the research questions and the dataset, a number of different segmentation procedures may be appropriate in second language research.

8.5.2. When to Make Coding Decisions?

Wherever possible, it is best to make decisions concerning how to code and how much to code prior to the data collection process—that is, when planning the study and preparing the protocol. By addressing coding concerns at the beginning—hopefully through a detailed pilot study—the actual collection of data can be fine tuned. For instance, if straightforward coding sheets are designed ahead of time based on research questions and variables, it may become obvious that the proposed data collection procedures cannot provide clear answers to the research questions. This may lead researchers to rework their plans for gathering data so that they can gather more information or different types of information from other sources. The best way to uncover and address such issues is by carrying out an adequate pilot study. This will allow for piloting not only of materials and methods, but also of coding and analysis. Designing coding sheets ahead of data collection and then testing them out in a pilot study is the most effective way to avoid potential problems with the data for the study.

8.6. CONCLUSION

Data coding is one of the most time-consuming and painstaking aspects involved in carrying out a second language research project. There are many decisions to be made, and it is important to remember that many of the processes involved in data coding can be thought through ahead of time and then pilot tested. These include the preparation of raw data for coding, transcription, the modification or creation of appropriate coding systems, and the plan for determining reliability. Careful coding is a key component of good research. In the next chapter we focus on quantitative data and, in particular, on what to do with data once they are coded and ready to be analyzed.