2-2004

Simplifying Qualitative Data Analysis Using General Purpose Software Tools

Nancy R. LaPelle
University of Massachusetts Medical School

Simplifying Qualitative Data Analysis
using General Purpose Software Tools

In Press: Field Methods, Sage Publications

Key Words:
Qualitative Analysis, Computer Assisted Data Analysis Systems (CAQDAS), Text Analysis

Abstract

This paper shows how clever but simple use of word processing functions can provide many features of special-purpose software designed for analyzing text. For many qualitative research projects, and for students who are learning computer-assisted analysis of text, the Microsoft Word functions outlined in this paper may be all that are required. Examples are given showing how Microsoft Word can be used for coding and retrieving, semi-automated coding and inspection, creating hierarchies of code categories via indexing, global editing of theme codes, coding of "face-sheet" data, exploring relationships between face-sheet codes and conceptual codes, quantifying the frequency of code instances, and annotating text. The techniques outlined can be used for analyzing and managing many kinds of data, including key informant interviews, focus groups, document and literature reviews, and open-ended survey questions.

Introduction

Multi-function programs for managing and analyzing text are widely available. For many qualitative research projects, however, the native functions of full-featured word processing programs can be used, with a little creativity, for many of the functions of dedicated QDA (qualitative data analysis) software. Ryan, for example (this issue), shows how coding and retrieval can be done using Microsoft Word macros. In this paper, I show how Microsoft Word can be used to perform these and other basic qualitative analysis functions.

Some researchers have been skeptical about using word processors for doing qualitative data analysis (Richards & Richards, 1994; Seale, 2002), particularly in regard to: automating the retrieval of similarly coded passages; handling large numbers of codes or many references from codes to text; conceptualizing about relationships between codes; and capturing data that may not be part of the texts themselves but rather are facts about the study informants, documents or organizations under study ("face-sheet data"). Also, using the macro language built into programs like Microsoft Word require programming skills that are beyond the capacity of most users of word processors.

Through trial, error, and necessity, I discovered that built-in functions of Microsoft Word (functions that do not require programming skill) serve admirably for many qualitative research projects. In fact, I have found that it is often preferable to use Microsoft Word to perform many basic QDA functions. Of course, dedicated QDA software excels in doing complex Boolean searches and in visualizing data, but these functions are not always needed for research projects.

I have used Microsoft Word to analyze text from process evaluations, case studies, key informant interviews, focus groups, and open-ended survey questions, among other sources of data. I use Word functions such as Table, Table sort, Insert file, Find/Replace, and Insert comment to do to this work. Projects have ranged in size from short simple tasks to complex multi-year research endeavors that involved over 200 interviews, over 2,000 pages of transcribed text and over 200 codes.

Using Word Tables for Coding and Retrieval of Interview Data

Miles and Huberman (1994) have shown that table structures are powerful tools for data analysis. When using Microsoft Word to support QDA, the Microsoft Word Table structure acts as a database that can be:

1) used to format informant, document or write-in survey data in a table structure for analysis
2) modified for coding purposes by adding rows and additional sort key columns to the table structure to accommodate coding
3) merged with tables of data for additional informants, focus groups, etc.
4) searched using the "Find" function for keywords or codes

5) sorted in a variety of ways (e.g., by theme code, by utterance number, by informant or focus group number, by gender, by informant role, by organization type, by question number, etc.) using the Word "Table Sorting" function
6) edited using the standard Word editing functions or the "Replace" function for global changes.

Dedicated QDA programs, like NUD*IST and Atlas/ti, provide relational database structures that store text, codes, face-sheet characteristics, memos/notes, and information about the linkages between these. If you tried to keep all this information in a single non-relational database or table, you would wind up with lots of duplication and the database/table would soon become unwieldy. The approach I will describe simplifies the tabular structures and concentrates on a smaller number of key relationships – namely linking text instances of predefined theme categories with theme codes and face-sheet codes. While this does not duplicate the power of a relational database, I have found that it is sufficient for many projects

The process for using Microsoft Word for coding and retrieval of qualitative data involves seven steps:

1) Format the data into data tables including participant ID information and utterance sequence numbers.
2) Develop a theme codebook in tabular format to define linkages between numeric codes and theme categories. Logically organize the codebook based on your framework or report outline.
3) Determine face-sheet data categories on which retrieval will be done and add columns to the data tables to accommodate coding for these.
4) Do the thematic coding in the theme code column modifying the table as needed to handle text that should be coded with multiple themes.
5) Sort the data by desired face-sheet data and theme code categories to look for patterns.
6) Validate the coding within a data table, correct and re-sort.
7) Merge appropriate data tables and validate coding across data tables. (Optional)

Step 1: Formatting Interview Data Into Tables

Most QDA programs require the data to be specially formatted before analysis. For example, many programs require that text be converted into ASCII format. NUD*IST and Atlas/ti suggest reducing the visible text to a 4- to 5-inch newspaper sized column to make coding easier. In order to use the Microsoft Word Table functions to process the data, there are also data formatting requirements.

Using the "Insert Table" selection on the Table menu on the main Word toolbar allows for the creation of a table into which data can be transcribed. At least a four-column table is required at the transcription stage and additional columns can be added later as needed. In this four-column table, each separate response of each speaker is entered into a new row of the table. Key informant or focus group participant response or utterance rows would be interspersed with interviewer questions in separate rows.

The data table excerpt below in Table 1 is from a focus group with school nurses interested in preventing teens from smoking. Although the order of the columns is not significant, the first column in this example is a unique speaker ID, and the second column will be used for the categorical coding or indexing. The actual utterances of the interviewer and participants are in column three. A chronological sequence number is entered in column 4 for each utterance at every speaker change. (This sequence number is important as it allows you to return to the original sequence of utterances should you ever sort the table based on other columns.) The moderator questions are in bold typeface to make them more easily visible. Transcribing in this tabular format allows the researcher to move directly into using Word for analysis. After transcription in the tabular format, the theme code column would be filled in by the analyst (Step 4) once the coding scheme has been designed (Step 2).

Table 1. Data Table Excerpt from Focus Group With School Health Nurses

Partici-pant Name	**Theme Code**	**Moderator Question/ Participant Response**	**Sequence #**
Moder-ator		**In order to get to know who is in the room, I'd like each of you to give a one minute introduction that includes your name, your school and any smoking cessation activities going on in your school.**	**1**
Pam		I'm Pam from "A" High School. I don't work directly with them in smoking cessation programs. I can do referrals to physicians for smoking cessation. And I do teach a health class as part of my nursing duties, so they get a lot of smoking information, in that respect.	2
Gerry		My name is Gerry and I'm at "B" High School, and the past two years I have done the Tobacco Education Program. We have offered the kids other programs this year with the local hospital that has had some smoking cessation programs and there's been hypnotists with this program, and we've been working very hard at it.	3
Paula		My name is Paula. I'm from "C" High School. We don't have a formal program. We try to reach out to the kids, on the Great American Smokeout Day; we do a big thing with that. We have for several years.	4

It is important to create a standard table template with consistent column widths if you plan to merge data tables from multiple informants at any stage in your analysis so the columns will line up after merging for sorting by theme code or other sort keys you have defined.

If an interview transcript has already been captured in an unformatted Word document, it can be easily converted into tabular format using the "Convert Text to Table" option on the Table menu on the main toolbar, as long as the text wraps from line to line and there are no intervening blank lines or hard returns (paragraph marks) within a single speaker's utterance. There should be hard

returns (paragraph marks) only at the very end of each speaker's utterance when the next speaker is about to begin speaking. When using the "Convert Text to Table" option, one needs only to specify "separate text at paragraphs" to create a one-row entry in the table for each new speaker. When doing the conversion, it is best to specify a single column table and then add the additional columns after the conversion. However, to avoid having to do the conversion, I generally have taped interviews transcribed directly into the tabular format I have described.

Step 2: Develop a Theme Codebook

Codebook development is treated in detail by many authors (Crabtree & Miller, 1992; Dey, 1993; Willms et al, 1990; MacQueen et al., 1998; Miles & Huberman, 1994; Araujo, 1995) and will not be covered in depth here. In preparation for analysis, a theme codebook is created by reading a representative sample of interviews and noting the themes that seem to reoccur or that have some significance to the study. The codebook should contain a definition of each major theme and each sub-theme within that major theme. The codebook also assigns numerical codes to the *in vivo* or constructed textual theme categories being defined. These numerical codes will be used for later sorting of text data by theme code. To ensure predictability of sorting, decimal numeric codes are used for the actual theme coding and a numeric sort is done on theme codes in the data tables during analysis.

To build a codebook, I use a separate table in Word (see Table 2). The codebook assigns numeric codes to theme categories and may also contain criteria for inclusion/exclusion of text instances and examples of these. It is up to the analyst to organize these codes in a logical way while creating a codebook table. Careful design of the codebook table and indexing structure can facilitate both analysis and reporting later on, since data will be sorted in the order that the numerical codes have been defined. If the theme codes are organized logically in the order of the outline designed for the report, both analysis and reporting can occur as the analyst proceeds sequentially through the sorted data table.

It is relatively easy to create hierarchies of related concepts and categories by designing the numerical coding or indexing appropriately. The type of indexing described in this paper is similar to that used by early versions of NUD*IST for this purpose.

Table 2. Three Level Codebook Table Excerpt

Level			
1	**2**	**3**	**Theme**
6.000			**Health site no longer offering stand-alone tobacco treatment services**
	6.05		No discussion about trying to sustain services
	6.10		Flavor of discussions about trying to sustain services
	6.15		Why was decision made to discontinue tobacco treatment services
	6.20		What is being done for smokers interested in quitting
		6.205	Screening only
		6.210	Physician counseling
		6.215	Counsel in-patients or those with serious health problems only
		6.220	Limited treatment as part of other regular visit

		6.225	Referral in-house to someone other than specialized tobacco treatment counselor

Designing the coding in this way will provide for the numerical sorting operation by theme code to order the coded response entries in the sorted data table in the same way the codebook has been ordered.

Step 3: Add Columns and Codes to Capture Face-sheet Data

In addition to developing thematic codes (Step 2), the analyst must also determine what other data are relevant for retrieval and then add additional columns for these to the table structure. For example, particular face-sheet categories not occurring in the data, such as gender, organizational affiliation, role within groups of interest, etc., may be of particular interest to the study. The simplest approach is to define one or two major retrieval categories in addition to the theme code and add additional columns for these to the table. These additional columns are especially helpful both to identify the source of the text and for sorting if the analysis will involve merging data tables for multiple respondents or focus groups prior to doing the retrieval. Some of these face-sheet categories can be combined in a single column as will be discussed later when I introduce some other features of Microsoft Word that are helpful in analyzing text in Step 5.

The face sheet columns can be used as sort keys in conjunction with the theme codes by which to group and retrieve data. Alpha codes can also be used for this face-sheet data. The data in Tables 3a-3c below came from key informant interviews used to identify success factors and barriers to launching publicly-funded tobacco dependence treatment services in community-based organizations. Three key informant interviews were done with staff in three different roles at each site.

In this example, three separate data table segments from the three key informants performing roles 1, 2, and 3 in organization "A" are shown in Tables 3a through 3c, respectively. Here we show only the segments of the data tables for theme code 1.55. Theme code 1.55 in column three was defined as the concept "community involvement in service definition in community health centers."

Table 3a. Data Table Excerpt for Informant 1 in Organization A

Org ID	Role ID	Theme Code	Interviewer Questions/Key Informant Responses	Sequence #
A	**1**	**1.55**	**Interviewer: Did you get any community involvement from your patients at any point to understand their needs for tobacco dependency treatment better, or find out what kinds of additional services they might want?**	**97**
A	1	1.55	That was done mainly by the outreach department within the clinic.	98

Table 3b. Data Table Excerpt for Informant 2 in Organization A

Org ID	Role ID	Theme Code	Interviewer Questions/Key Informant Responses	Sequence #
A	**2**	**1.55**	**Interviewer: Has there been any involvement from the community in terms of getting input from them to tailor the services -- sort of a needs assessment?**	**51**
A	2	1.55	We have not done that since I've been there, and I don't know frankly if it was done at any point. There is a lot of smoking in the community, we know that.	52

Table 3c. Data Table Excerpt for Informant 3 in Organization A

Org ID	Role ID	Theme Code	Interviewer Questions/Key Informant Responses	Sequence #
A	**3**	**1.55**	**Interviewer: Was there any community involvement in getting feedback relative to the way the services would be offered at any time?**	**73**
A	3	1.55	Yes. We -- of course we have a consumer board of directors. There have been focus groups that have been done. We have a very active outreach program.	74

Column one in Tables 3a-3c is used to indicate that these three data tables all belong to key informants from organization "A". The second column is used to indicate the organizational role performed by each informant. These numeric or alphabetic face sheet codes also have to be defined in the codebook.

Step 4: Coding Text Rows with One or with Multiple Theme Codes

Qualitative analysts have typically used two different types of coding approaches (Bernard, 1991 and 2002; Seidel & Kelle, 1995). In classic content analysis (e.g., Krippendorf, 1980; Weber, 1990), codes are assigned as values (categorical, numerical, or interval) to fixed units of analysis where one or more codes can be assigned to a unit but only in its entirety (e.g., paragraphs, answers to open-ended questions, turn-taking in a focus group). In grounded theory (e.g., Glaser & Strauss, 1967; Strauss & Corbin, 1990; Dey, 1993), a different approach has been used where codes refer to the process of tagging or marking contiguous blocks of text that can include variable units that range from simple phrases to text that extends across multiple pages within a corpus of material.

Coding as tags can apply to text that overlaps both completely as with the former coding approach or only partially. Using the technique I describe next, researchers can apply either coding approach depending on how they choose to use it. The examples illustrate coding as both

tagging and assigning value since: 1) using the theme code as a tag for segments of larger utterances, one can tag and retrieve the full text of variable units of data for selecting illustrative quotes to enhance reporting; and 2) one can also code occurrences of multiple interpretive theme categories within a fixed text unit via the numerical codes assigned in the code book by assigning multiple codes to the same text unit.

When all of the text in a row falls into a category you have defined, such as in Table 4 below, you can simply enter one theme code per row of the table. In this example (from a focus group to evaluate a stop-smoking intervention that included suggested dietary and physical activity changes to support quitting), code 4.03 or 4.030 was the moderator question about what worked well or needed improvement in the dietary change part of the intervention. The respondent theme codes simply related to what worked well (4.031) and what did not work so well (4.032) in this part of the multi-faceted intervention.

Table 4. Coding Text Segments with One Theme Only

Partici-Pant ID	Theme Code	Question/Response	Sequence #
Mode-rator	**4.030**	**How did you feel about the dietary information you received about ways to avoid gaining weight and stay healthy while quitting smoking?**	**329**
Laura	4.031	I think a lot of it was an education about things, like why weren't we using alternatives to milk because of the hormones in the dairy products. I thought it was very interesting because you don't tend to think about that and how it can affect you and using the soy products which at first we all were so disgusted to think of soy. It was soy everything. But you know most of it wasn't bad.	330
Justin	4.031	Well they kind of switched toward the end.	331
Nora	4.031	Well I gave feedback - please do something in real life that people are really going to use because. And then they did, they listened. I think they listened to our feedback.	332
Don	4.032	Broccoli. No more broccoli.	333
Laura	4.031	We did like the chili with the soy. I thought that was very good.	334
Don	4.031	I felt like I would never in a million years cook with tofu, never. I went home and cooked that tofu lasagna and it was really good. Being able to try it here - that was very helpful. I have changed my eating habits tremendously from this group. And I think that doing something different was giving me a tool that I didn't have. I found that to be very helpful.	335
Justin	4.031	Yeah, they challenged us to try different things. You're going to leave here and 6 nights a week you're going to have what you always have and one night a week I was looking forward to coming here and trying something odd. (laughing) And some of it was fantastic. I drink soy milk every day now because of that class. I'd still be walking by it in the supermarket saying, not that	336

		stuff, but it's delicious.	
Laura	4.032	I didn't care for the bean burgers (all laughing)	337
Mode-rator	**4.031**	**So it was helpful to be exposed to these different things although you might not choose to…**	**338**
Laura	4.031	Yes, the exposure was key. It made you think about what you ate on the following day.	339

In order to make sure the moderator question stays with and precedes the relevant responses after sorting, it is often best to code the moderator question in a way that ensures that this will happen, i.e., with the same code number as the first theme in the participant response or with the highest level relevant code category. Otherwise, the analyst may find that the sort leaves her or him with response data to an unknown question that wound up sorting somewhere else. This can easily happen when: 1) the informant digresses or decides to refine his/her answer to a previous question instead of answering the one just posed; or 2) the moderator questions are coded by the theme the moderator question reflects rather than the theme occurring in the response given by the informant. Examples of assigning multiple codes to a single utterance are given below in tables 5-7.

As part of the coding and analysis, the data table structure can be modified to assign multiple codes when multiple themes occur in a single participant response/utterance. This can be done by adding a table row using the Insert Row option on the Microsoft Word Table menu and either splitting or duplicating the text as we will see in the examples below. The investigator can split the text in a row entry at any point and is not bound by sentences or paragraph endings. The decision as to whether and where to split or block the text depends entirely on the investigator's perception of where one theme ends and another begins in the text and whether or not quantification of occurrences is to be done. Quantification imposes constraints to avoid proliferation of occurrences within a particular utterance. This is discussed further at the end of this section and later, in the section on "Other Useful Word Processing Functions" (i.e., Counting Instances of Themes).

Several focus groups were held to get reactions of men and women, ages 45–65, to a program for preventive health care. Participants were asked to react to a newly developed draft brochure they would receive in the mail urging them to be screened for colorectal cancer. The brochure mailing would be followed by a motivational phone call three months later to remind them of the brochure and again urge them to get screened.

Table 5 shows a coded excerpt from one of these focus groups. The moderator is asking specifically for participants' reactions to receiving a follow-up phone call three months after receiving the mailed brochure. In this example, Will's utterance (sequence # 552) falls into several code categories defined in the codebook. Code 8.05 relates to the time interval between the receipt of the brochure and the follow-up call; code 8.10 relates to suggestions that a written notice should be sent just prior to the call to remind brochure recipients to expect a call; code 8.15 relates to tips to increase the response rate to the call; and code 8.20 relates to overall feelings expressed about receiving such a call. Note that the theme codes are not overlapping in the response text in this example, and each theme occurs only once.

If part of the response/utterance relates to one theme and other parts relate specifically to other themes, the analyst can segment the response into multiple rows, as has been done in Table 5, by creating new rows in the table, copying the face sheet data from the original row into the new rows, and cutting and pasting the text that relates to each theme into the new rows. In Table 5, I have segmented Will's response 552 into response number 552.01, 552.02, and 552.03 in the rightmost column and have given each new row a different theme code in column 2.

Note that the decimal suffixes on the sequence number are a reminder to the analyst that this row was originally part of a longer utterance. I use .01 instead of .1 as the first suffix to accommodate narrative data tables where one utterance may last for several pages prior to coding (the suffix of .1 would allow only 9 segments). In this way it can be broken into up to 100 segments. If more segments are needed, .001 could be used as the suffix. Also note that the moderator text has been given the higher level theme code 8.00 to make sure that it will sort before any responses on the topic of the telephone intervention.

Table 5. Coded Focus Group Data Table Excerpt

Name	Theme Code	Moderator Question/Participant Response	Sequence #
Mode -rator	**8.00**	**This particular grant is testing an approach to motivating people to get screened for colon cancer. The first piece is to send them a brochure that you have just provided us with feedback on. The second piece of it is a follow-up phone call. The current game plan is three months after the brochure is sent to people, the phone call comes. How do you feel about having somebody call you up on the phone to talk about colon cancer screening?**	**551**
Will	8.15	Oh, just to get beyond the caller-ID screening, they have to use a University phone number, so that when it comes up on your Caller ID you don't say, 'I don't want to answer this call.'	552.01
Will	8.05	As far as the follow-up call, if it's three months it's going to be too long. If you call a month later, something might have already been planted in somebody's head from reading this. You call a month later, they might have done something about it by then. But if they haven't done something about it by then, then it might be the impetus just to push them into it.	552.02
Will	8.20	I don't know, that wouldn't be anything that would really bother me, to get a phone call after I received this. I mean, they call for every other blasted thing in the world these days. This one wouldn't be a bad one. This one is for my own good.	552.03

As long as each theme occurs only once in the utterance, segmentation like this will not affect quantification of occurrences within a single interview data table. If a theme occurs more than once in the same utterance separated by text that is an instance of a different theme, you will increase the count of occurrences unnecessarily if you segment the second occurrence into its

own additional row. Instead, it may be better to put all text relating to the same theme from a particular utterance in one row and to signal that there was intervening text with "[….]." A row would be added for the differently coded intervening text. Reverse brackets "]" and "[" could be used at the beginning and end, respectively, of the intervening text in its own row to indicate visibly that the intervening text represented a different theme from the preceding and subsequent text.

If the whole response seems relevant to more than one theme as in Table 6 below, one can simply duplicate the row and give the second instance of the row a different theme code in the theme code column. In the data table excerpt in Table 6 from key informant interviews to identify barriers and success factors to implementing successful tobacco treatment programs, the response falls into two success factor theme categories:

2.06 Program Staffing
2.10 Internal Administrative Systems.

But the response text, in this instance, is not as easily segmented, so one way to handle this is to duplicate the entire response with two different theme codes. To keep track of text that has been duplicated, it is a good idea to change the font of all the copies of the duplicated text. Italics are used below. Additionally, it can be given a suffix to indicate how many times it has been duplicated. In this case, two copies of the data will occur in the data table and the suffix .0002 or .00002 can be used to flag this visually as well. Note that both rows get the sequence number of "72" since the row has not been split.

Table 6. Key Informant Data Table Excerpt

Parti-cipant ID	**Theme code**	**Question/Response**	**Sequence #**
1.3	**2.06**	**Interviewer: Were there other things in place that contributed to the success, in terms of administrative procedures, or ...?**	**71**
1.3	*2.06*	*Well, I think that Jesse [i.e., the program manager] has the skills in terms of being able to create a tracking mechanism and evaluation process that allow us to actually track who we're seeing and how we're performing. That I think is really very important. We've been really committed always to evaluate the process.*	*72.0002*
1.3	*2.10*	*Well, I think that Jesse [i.e., the program manager] has the skills in terms of being able to create a tracking mechanism and evaluation process that allow us to actually track who we're seeing and how we're performing. That I think is really very important. We've been really committed always to evaluate the process.*	*72.0002*

In the process of assigning codes to units of text, researchers often find that they need to add notes that help fill in the gaps about what was being said. In such cases, a set of conventions is useful. Brackets, as in Table 6, mean that text has been left out or inserted by the analyst during coding. Another example of when this kind of explanatory insertion indicated by brackets is helpful is if the text contains a pronoun or backward reference (e.g., he, she, it, this, that, etc.) that may get sorted away from its antecedent and lose its meaning. Another situation where bracketing is helpful is when adjoining text that has been segmented into a different theme code is needed to provide the context for the other half of the segmented text but is not actually part of the theme.

An example of this is below, beginning with the uncoded text in Table 7a.

Table 7a. Segmenting Overlapping Themes

Partici-Pant ID	Theme Code	Text	Sequence #
Howard		John's anger spiraled out of control. His eyes blazed and his jaw tensed. Then he yelled at his brother for burning the steaks. This wasn't the first time his brother had screwed up.	26

If we were coding this for the themes of 7.00 relating to expressions of anger and 12.00 relating to assessments of competency, we could either duplicate the whole utterance and assign it two codes as in Table 6, or we might segment and code it as in Table 7b where the brackets indicate *in vivo* context for the segmented text.

Table 7b. Using brackets to provide *in vivo* context for segmented text

Partici-Pant ID	Theme Code	Text	Sequence #
Howard	7.00	John's anger spiraled out of control. His eyes blazed and his jaw tensed. Then he yelled at his brother [for burning the steaks.]	26.01
Howard	12.00	[Then he yelled at his brother] for burning the steaks. This wasn't the first time his brother had screwed up.	26.02

One might also consider another way of doing this as below in Table 7c.

Table 7c. Using Text Segmentation and Duplication to Deal with Overlapping Themes

Partici-Pant ID	Theme Code	Text	Sequence #
Howard	7.00	John's anger spiraled out of control. His eyes blazed and his jaw tensed.	26.01
Howard	7.00	*Then he yelled at his brother for burning the steaks.*	26.0202
Howard	12.00	*Then he yelled at his brother for burning the steaks.*	26.0202
Howard	12.00	This wasn't the first time his brother had screwed up.	26.03

In the first solution in Table 7b, the text has been segmented into only two pieces. It has been segmented into three pieces in the second solution and the sequence number suffixes indicate this in each case. In the second solution in Table 7c, the sequence number suffixes also indicate that the second segment has multiple themes, has been duplicated and occurs in the data twice. However, if segmenting is done this way, as in Table 7c, this will add unnecessarily to the count of theme occurrences if quantification is being done across short responses in a larger transcript formatted in a data table.

To avoid this, it is preferable to use either the duplication techniques as in Table 6 or the segmenting technique in Table 7b. If a particular theme occurs multiple times in a longer text that is basically one text row, such as in a narrative or document, however, it may be advantageous to count each time it occurs in a single text entry, and the row may be segmented many times with many different theme codes. How the segmenting and quantification is to be done must be considered up front and designed into the way coding is done for the particular data set.

Assigning thematic codes to text units is hard work, repetitious, sometimes boring, and prone to errors. One way I have found to overcome some of these problems is to look for-occurrences of keywords relevant to a particular theme code that can be inspected in context and coded appropriately. For example, if key informants or participants at a focus group have used key words that are related to one of the constructs of significance for the study, coding can be simplified by searching the data table for those key words and assigning theme coding as they are found.

To do this, the Find option on the Microsoft Word Edit menu can be used. Word will stop the search automatically as each occurrence is found to allow the analyst to inspect the context and decide if this occurrence is truly an instance of the code category in question.

For example, if an analyst searched Table 1 for "smoking cessation programs," using the Word Find function on the Edit menu, the analyst would be directed sequentially to the responses given by Pam and Gerry, who both mentioned these keywords. If the analyst wanted to code for schools that were offering smoking cessation programs in the school, however, the analyst would have to read the context carefully to find out that these school nurses are actually saying that they must refer their youthful smokers out to cessation programs in the community. Once this semi-automated coding has been done for *all* relevant keywords, the analyst can then read the entire data table to code other relevant data that did not include any anticipated key words.

Step 5: Sorting Data Tables and Finding Patterns

Once you have coded the text for themes and assigned appropriate face-sheet codes (i.e., participant characteristics) to each text unit, then you can move on to doing more complex analysis by searching and sorting your data based on different analytic criteria. You can do relatively sophisticated searches by using Word's Table/Sort function. For example, you can sort by thematic code, participant characteristic, or sequence number. Sorting on combinations of

these face sheet codes, along with a specific theme code from the text, will group all the data for the combined codes together for further review and analysis.

Sorting is done using the numeric theme code column as the primary key, face-sheet data column(s) such as the participant ID column(s) as the mid-level sort key(s), and the numeric sequence number as the lowest level sort key. The sequence number is always the lowest level sort key to keep the sorted data in chronological order. Select ascending or descending sort depending on the order in which you want your results to appear, and make sure that you have selected numeric sort keys, except when you have defined textual face-sheet data codes.

The Word Table function can accommodate the use of only three sort keys in one sorting operation so the face-sheet variables of interest can sometimes be combined, if desired, to stay within the three sort-key limit to avoid doing a two-pass sort. This has been done in Table 6 where the participant ID column combines the organization number and a staff role number (i.e., 1.3, where this refers to organization 1 and role 3). While combining face-sheet codes can eliminate the need for additional columns and additional sort passes, it constricts the flexibility of sorting options. Table 6 can only be sorted by role ID within organization ID and not vice versa.

In Table 8, the transcripts for four informants in organizations A, B and C have been merged and sorted and we see an excerpt from the resulting data table file for code 1.55. Two columns were used for the organization ID and role ID. If the analyst wanted to look for patterns across roles within organizations, data tables for all three roles and all three organizations A, B, and C, could be merged. The merged file could then be sorted using the text theme code as the primary key, the organizational code as the secondary key, the role code as the tertiary key, and the sequence number as the fourth sort key.

First, the two data tables for roles 1 and 2 for organization A were merged (see steps 6 and 7 below), and then data tables for organizations B/role 2 and C/role 1 were merged with that for organization A. The resulting file was then subjected to a two-pass sort to accommodate the four sort keys. The first sort is simply by numerical sequence number. The second pass was sorted by numerical theme code, textual organization ID code and numerical role ID code. The general rule is to sort first by lower level sort keys 4-5-6 (depending on how many sort keys are being used), and then by higher level keys 1-2-3 on the second pass.

Table 8. Excerpt from Merged/Sorted Data Table for Four Key Informants in Three Organizations – Sorted by Theme Code, Organization ID, Role ID and Sequence #

Org ID	Role ID	Theme Code	Question/Response	Sequence #
A	**1**	**1.55**	**Interviewer: Did you get any community involvement from your clients at any point to understand their needs better, or find out what kinds of additional services they might want?**	**97**
A	1	1.55	That was done mainly by the outreach department within the clinic. The only thing I did in that respect was to talk to them	98

			once a month, the outreach group -- but other than that I limit myself mainly to the health center, because you see the health center -- with over 150 employees -- has a lot of clients.	
A	**2**	**1.55**	**Interviewer: Has there been any involvement from the community in terms of getting input from them to tailor the services -- sort of a needs assessment?**	**51**
A	2	1.55	We have not done that since I've been there, and I don't know frankly if it was done at any point. We do know -- we know that in this population that we serve, the percentage of smokers is higher than in the greater community. There is a lot of smoking in the community, we know that.	52
B	**2**	**1.55**	**Interviewer: Did you get any feedback from smokers relative to what their needs were that might have provided input into how the services were delivered differently ?**	**65**
B	2	1.55	Well we did do evaluations for all our programs at the end of the program. And we had a really comprehensive, uh, program. And we certainly did listen to peoples' requests.	66
C	**1**	**1.55**	**Interviewer: To what extent has community involvement driven the evolution of services?**	**41**
C	1	1.55	Community involvement. Good question. Not sure I know how to answer that.	42

If the analyst were instead looking for patterns within roles regardless of organizational affiliation, he or she would reverse the secondary and tertiary sort keys for the sort. Adding the two extra columns for the two face sheet variables provided this flexibility in sorting options. This additional sort could not have been done using the combined face sheet data key in Table 6.

One function Word does not provide is the ability to concurrently search for relationships between several different themes that occur in the text of the responses themselves, e.g., "smoking cessation programs" and "hypnotists" in the example in Table 1. However, it is quite easy to do these searches manually, once the data are sorted by theme code, by simply looking through the sorted data table for the two codes of interest. Ryan (this issue) describes a different approach for doing simple Boolean searches such as those described here by using Microsoft Word macro techniques.

Steps 6 and 7: Code Validation/Correction and Merging of Data Tables

Code validation is easy using this data table sorting technique both within and across interview data tables, even with multiple coders.

Code validation within one interview data Table. Once the data table for a single interview has been sorted by theme code and sequence number, the analyst reads all text segments for each code and decides whether text segments are all instances of a particular category or if corrections are needed. Any erroneous codes in the data table can be corrected and then the data table can be resorted.

Merging data tables. To do code validation across interviews, data tables from all interviews or some subset of them appropriate to one's study can be merged (e.g., all those for a particular gender, organization, type. etc.). Note that sorting after the merging operation will only work easily if the same template is used for all the data tables in a study so that all the columns line up. To merge data tables, simply insert the additional data table files one at a time at the end of the previous data table file using the Word "File" selection on the Insert drop down menu, deleting any intervening space between tables. Merging many files can get unwieldy depending on the memory and computing power of your computer so it may be prudent to merge and sort only subsets that are related in some way for analysis. I have merged up to twenty data tables, but generally do this in smaller subsets.

Code validation/correction across interview data tables. Once merging has occurred, code validation across transcripts can be done. To do this, sort by theme code, face-sheet codes of interest such as participant ID column(s), and sequence #. Once the data tables for all appropriate interviews have been merged and sorted, the analyst again reads all text segments for each code and decides whether text segments are all instances of a particular category or if corrections are needed. Once corrections are made in the merged/sorted data table file, it can be resorted.

Other Useful Word Processor Functions

Making global coding changes. Sometimes an analyst may want to change all entries coded with a particular theme code to another theme code or to a different higher level theme code grouping or merge them with another code. Global coding changes can be done easily by highlighting the theme code column and then using the Microsoft Word Replace All option on the Edit menu to specify the old theme code and replacement theme code. Re-sorting the file after the global replacement will then regroup these themes so the data table is once again in ascending numerical order by revised theme codes. It is important to make sure the codebook reflects any global code changes such as these.

Memoing. Word also provides a feature that can be used to insert comments or analytic memos, as they are known to grounded theorists, with the individual or merged and/or sorted data tables if this is desired. This function can be used to note emerging ideas or definitions of codes. Using the Insert drop down menu, select the "Comments" option that will place a marker in the text and open another window where the comment can be entered. Alternatively, one can use the View menu. View Toolbars will allow the analyst to select the "reviewing" toolbar. This also provides the capability to insert, review and edit comments related to particular segments of text.

The analyst can decide if this feature is more useful than keeping such notes and memos elsewhere. This feature is designed for adding review comments to documents so it is not as flexible as it could be for analytic memos; however, one can print all these comments separately via the Print dialog box by selecting "Comments" in the "Print What" box or select and Copy them to another document that could serve as the beginning of a report.

Counting instances of themes. Many QDA programs provide the capability to automatically tally up the number of instances within a theme category and the number of informants generating these instances. One can do a rudimentary frequency count using Word by highlighting the theme code column, selecting the Replace function on the Edit menu. If one is counting all instances of code 2.031, for example, one can automate the counting by using Replace All and specifying the code as both the text to be found and the replacement text. Word will report how many replacements were made giving the number of code instances. However, after the data table has been sorted, it is almost as easy to simply count these manually since they are all grouped together, unless there are many data tables that have been merged and many instances of each theme. Splitting or duplicating data table rows for coding multiple themes in one utterance may artificially increase the counts for a particular theme code for a particular utterance if the analyst does this without quantification in mind. Going back to the alternate solutions in Table 7b and Table 7c, one can see that the counts of occurrences would be different in the two solutions. Counting the number of people generating code instances must be done manually, but, again, this is facilitated by having all instances of the code grouped together as a result of the sort.

Note that when counting all instances of a particular theme code, the comments of the moderator or interviewer that have been assigned that theme code will also be counted if they have not been assigned a different (but related numeric code). To avoid this, a theme code such as 2.03 could be modified so that all interviewer questions are coded 2.030 and all interviewee responses are coded 2.031.

Handling multiple documents via windowing. A feature that can be particularly useful during analysis is the multiple windowing feature of Word. If one is analyzing a merged and/or sorted data table and wants to see the original context of a particular text segment, one can open the original data table and search for this segment and easily see the surrounding context. Switching back and forth between the open windows is accomplished using the pull-down Window menu on the main toolbar.

Open-ended survey questions. Open-ended survey question data can be analyzed easily using similar techniques. The analyst or transcriber simply creates a four-column table such as the one below in Table 9 with columns for the question number, the respondent ID, the question or response text, and the theme codes. Once the data have been entered, codes can be developed that relate either to the question number or to conceptual themes and these can be entered in the code column. It is often desirable to use conceptual theme codes rather than question number codes as the highest level (whole number) coding category since survey respondents may well write in data that is relevant to other open-ended questions or that refer back to these, and coding by question number will not allow sorting this type of data into the appropriate thematic category.

Note that survey question responses can be transcribed in any order in the data table since they can later be ordered by question number or by theme code via a simple sorting operation. Zero has been used for the respondent ID of the questions to ensure they sort before all the responses. In this format, the data table can be sorted by either theme code or question number for further analysis.

Table 9. Survey Write-In Data Table for Process Evaluation Feedback

Question number	Respondent ID	Write-in Response	Theme Code
14	**0**	**What worked best in implementing this initiative?**	
14	23	Everyone who stayed involved to the end collaborated and supported one another despite differences of opinion. However, we lost one key person early on.	
19	**0**	**What worked least well in implementing this initiative?**	
19	78	One key person was not really committed and dropped out of the effort early on.	

Conclusion

For many, though not all, data management and analysis functions, Microsoft Word can be used as QDA software. There are clearly some instances where dedicated QDA software is superior, e.g., in handling visual data and in doing complex Boolean searches across text-based categories. However, for those who do not need these features, the approach I have described provides an inexpensive path with a short learning curve to semiautomation of many QDA tasks.

References

Araujo, L. (1995). "Designing and refining hierarchical coding frames," in Computer-aided qualitative data analysis: Theory, methods and practice. Edited by U. Kelle, pp, 96-104. London: Sage Publications.

Bernard, H.R. (1991). About text management and computers. Cultural Anthropology Methods Journal 3:1-4, 7, 12.

Bernard, H.R. (2002). Research methods in anthroplogy: Qualitative and quantitative approaches, 3rd edition. Thousand Oaks, CA: Sage Publications.

Crabtree, B.F., and W.L. Miller (1992). Doing qualitative research. Newbury Park, CA: Sage Publications.

Dey, I. (1993). Qualitative data analysis: A user friendly guide for social scientists. London: Routledge and Kegan Paul.

Glaser, B.G. and A. Strauss (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago: Aldine.

Krippendorf, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills, CA: Sage Publications.

MacQueen, K., E. McLellan, K. Kay, and B. Milstein (1998). Code book development for team-based qualitative analysis. CAM Journal 10:31-36.

Miles, M. B. and A. M. Huberman (1994). Qualitative data analysis: an expanded sourcebook, 2nd edition. Thousand Oaks, CA: Sage Publications.

Richards, T.J. and Richards, L (1994). "Using computers in qualitative research" in Handbook of Qualitative Research, Denzin, N.K. & Lincoln, Y.S. (eds.) pp 273-285. Sage Publications: Thousand Oaks, CA.

Seale, Clive F. (2002). "Computer assisted analysis of qualitative data" in Handbook of Interview Research, Gubrium, J.F. and Holstein, J.A., editors. Sage Publications: Thousand Oaks, CA.

Seidel, J. and U. Kelle (1995). "Different functions of coding in the analysis of textual data," in Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice. Edited by U. Kelle, pp. 52-61. London, Sage Publications.

Strauss, A. & Corbin, J. (1994). "Grounded theory methodology" in Handbook of Qualitative Research, Denzin, N.K. & Lincoln, Y.S. (eds.) pp 273-285. Sage Publications: Thousand Oaks, CA.

Weber, R. P. (1990). Basic content analysis, 2nd edition. Newbury Park, CA: Sage Publications.

Willms, D. G., D. W. Taylor, et al (1990). A systematic approach for using qualitative methods in primary prevention research. Medical Anthropology Quarterly 4:391-409.