

# Analyzing Large Free-Response Qualitative Data Sets – A Novel Quantitative-Qualitative Hybrid Approach

Jennifer Light, Ken Yasuhara  
jlight@lcsc.edu, yasuhara@cs.washington.edu

**Abstract** - Qualitative analysis tends to be unwieldy for large data sets yet is an indispensable tool for understanding how and why phenomena occur. Consequently, the goal of this study was to develop a method that is credible yet economical for large, specific, qualitative data sets. The strength of our hybrid, qualitative-quantitative method comes from using automated text analysis techniques to focus resource-intensive coding efforts on a small, carefully selected subset of data. This paper details the hybrid method as applied to a previously analyzed set of free-response data and argues for the method's validity by comparing results from the hybrid analysis with the previous traditional qualitatively analyzed method. With this data set, the hybrid method yielded comparable results with substantially less manual coding and in less than a third of the time required for the original analysis method.

This hybrid analysis provides a more economical alternative for a “coarse-cut” qualitative analysis and observation of long-term trends, providing insight to practitioners, assessors, and researchers ranging from individual course evaluations to large-scale studies. Short, focused, open-ended survey questions are good candidates for this type of analysis.

*Index Terms* – Hybrid qualitative-quantitative, Large data sets, Method development

## BACKGROUND

The idea for a new kind of analysis—one that allowed discovery of a new idea, that would reduce researcher bias, that could be done by someone with little qualitative experience, and that would not take a large investment in time—was the impetus for developing a hybrid, qualitative-quantitative method for analyzing large sets of qualitative data. The generalized method steps are shown in Figure 1. This paper outlines Steps 1 and 2 then follows the “check” part of Step 3 for corroborating results from a previous analysis. When using this method for analysis (versus method corroboration) the “check” part would not be followed.

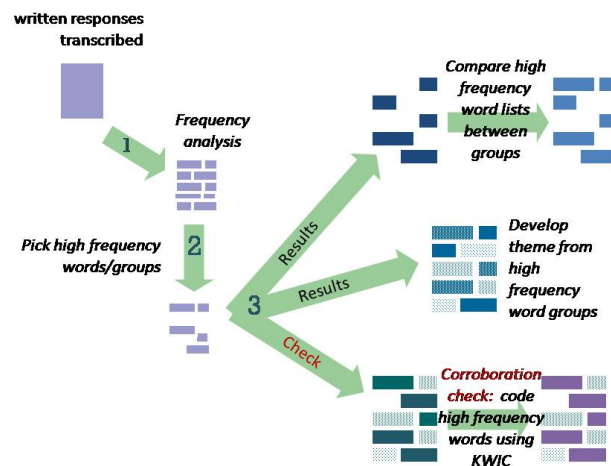


FIGURE 1  
GENERALIZED METHOD STEPS.

Although word frequency and qualitative coding is not a new idea, combining them in this way is a novel approach. Generally, high-frequency word analysis is done as a part of a content analysis. Content analysis is the process of reducing text material to manageable, relevant bits of information coded into categories [1]. This type of analysis tends to be valued for its reduction in research bias, since neither the researcher nor respondents are aware of the message being analyzed [1]. Content analysis, specifically word frequency, has been used for understanding cultural changes and phenomena by analyzing words of songs, speeches, and written communications, for example.

Qualitative research, defined by Strauss & Corbin [2], is “any kind of research that produces findings not arrived at by means of statistical procedures or other means of quantification.” Traditionally, qualitative types of research include interviews and observations but also can include document analysis and videotapes. At its heart, qualitative research employs a systematic set of procedures to develop and inductively derive theory about a phenomenon [2]. In essence, there is no *a priori* theory; rather the findings are derived from the data from which a theory is constructed.

Quantitative and qualitative data methods are often combined to produce a mixed-methods result, potentially offering the best of both worlds. Ideas and theories that may have been overlooked in a pure quantitative study are retained, while the phenomenon occurs frequently enough to substantiate claims of generalizability. These ideas were the basis for the framework of this hybrid approach. The intent for the larger study (of which the data set used for developing this method is part) was to understand the changes of engineering students as they go through their academic careers [3]. Using content analysis (in the form of word frequency) paired with qualitative analysis appeared to be a logical methodological pathway for understanding cultural phenomena (engineering school) and discovering changes experienced by engineering students.

The data set consists of transcribed free responses of 124 first-year engineering students at four different types of higher education institutions in the U.S. Specifically, during a timed session, students responded in writing to the question, “Over the summer the Midwest experienced massive flooding of the Mississippi River. What factors would you take into account in designing a retaining wall system for the Mississippi?”

So how good is this new hybrid method? As with most qualitative research, the results should be relevant, generalizable, verifiable, and consistent [2]. An earlier analysis of the same data set used a traditional qualitative approach [4]. These results were compared with a subset of coded, high-frequency words/word groups using the new hybrid method.

#### *Background for previous comparison study*

In the previous study [4], the 124 transcripts were divided into 1418 “thought units” or *segments* and coded. Two researchers independently coded each segment, achieving 80% agreement and negotiating disagreements to consensus. As illustrated in Figure 2, each segment was coded on two dimensions: physical location and frame of reference. From these two-dimension pairs, each segment was categorized as oriented toward either design detail or design context, as shown in Figure 2.

### METHODOLOGY

The process to develop the new method is broken down into eight steps. The first six steps describe the process for developing the list of high frequency words/word groups, and the last two steps are for method corroboration with the

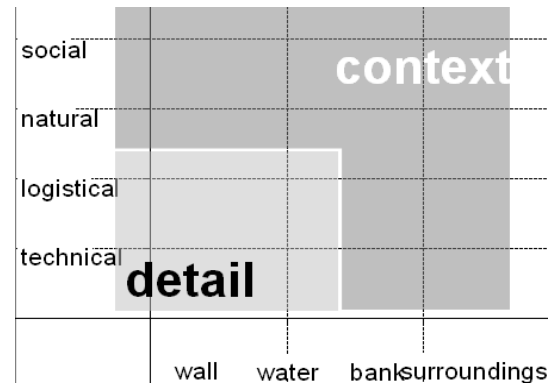


FIGURE 2

TWO-DIMENSIONAL CODING SCHEME, FROM KILGORE ET AL. [4]

earlier study that used traditional qualitative coding to analyze the data. The steps are 1) transcription and transcription cleaning; 2) arrangement of transcribed files into corpora (groups of files) analyzed for frequency; 3) determining and deleting stop words; 4) grouping similar words into “word groups”; 5) reanalyzing for frequency using word groups; 6) determining which word groups were high frequency (i.e., the cutoff for what words were considered high frequency); 7) coding high-frequency word groups using the original study codebook; and 8) comparing coded results to the original study results.

#### 1. *Transcribing and cleaning the data*

Student responses to the Midwest floods question were transcribed into Word files, maintaining the structure, spelling, and other markings of the original paper. To prepare the files for word frequency analysis, all misspelled words were corrected, non-letter symbols were changed (e.g., “H<sub>2</sub>O” was converted to “water”, “\$\$” was converted to “money”, → symbols were removed), and transcriber notes such as “unreadable” were removed.

#### 2. *Frequency analysis*

After cleaning the transcript files, they were grouped by institution and gender in to corpora. For each corpus, a list of high-frequency words was produced using the “word forms” feature in TextStat© [5], free software for word frequency analysis. Figure 3 shows a screen shot of a partial word frequency list from one corpus.



6. *Cutoff for determining a high frequency word/word group*

Occurrences of high-frequency words were counted on a per-participant basis. A cutoff of 0.5 word/word group occurrences per participant or higher was chosen to denote a high-frequency word/word group. Dividing by the number of participants, rather than the total number of words across participants' transcripts, helped to minimize the effect of variation in writing style. For example, in one corpus consisting of transcripts of 44 participants, the word "retaining" occurred 36 times. The per-participant average frequency for "retaining" was  $36/44 = 0.818$ ; consequently, "retaining" was considered a high-frequency word, since its average frequency was above 0.5 occurrences per participant. The usefulness of dividing by participant count, rather than word count, is illustrated by considering the alternative with the above example. There were 4515 total words in the 44-transcript corpus. Dividing by the total word count yields a word-level frequency for "retaining" of  $36/4515 = 0.00797 \times 100$  or 0.797%. Because some respondents used essay-style writing in their responses (tending toward more words) and some used list-style responses, per-participant frequency was chosen over word-level frequency.

7. *KWIC coding with original study codebook*

Occurrences of four high-frequency word/word groups with respect to context and detail were aggregated and analyzed using the concordance feature in TextStat and coded for corroboration with the previous study. The four word/word groups used were wall, water, flood, and effect/affect. Using a window of 70 characters before and 70 characters after, occurrences of the important words were situated in context and were coded based on that information. Only one idea was used per important word occurrence. If more than one idea was close to "wall," then the idea closest to "wall" in the sentence was used. For example, if a transcript contained the phrase, "wall cost and maintenance," then "cost" was coded as it was closest to the important word, "wall." If the important word was used more than once in the same sentence and was referring to the same theme category, the idea was counted twice. An example of this is the phrase, "the length and thickness of the wall, the height of the wall, and location of the wall...", which would correspond to three usages of the word "wall" coded in the two-dimensional coding scheme as wall-technical, wall-technical again, and wall-logistical for the third "wall" usage.

8. *Comparing results*

Proportional amounts (number of segments assigned a particular code pair divided by the total number of segments) of the two-dimensional coding for context and detail, along with context-detail percentages, were compared. Results

from coding 20% of the transcripts using the hybrid method were graphed using the same technique developed by Kilgore et al. [4] for visual comparison to the original study. The hybrid subset of transcripts maintained the same proportions of male and female respondents. This was done because the previous research indicated a significant difference in the detail-context orientation by gender. The graphs and proportional percentages are found in the results section.

RESULTS

The results from two-dimensional coding of the data set from the previous study and the new hybrid analysis are shown in Figures 5 and 6, respectively. The numbers in both graphs are percentages of the total data sets used for each analysis. In Figure 5 the number shown for each two-dimensional pair is the percentage of responses for that code divided by the total number of segments. For example, there were 1418 coded segments for the entire original study; of those, 240 segments were coded wall-technical. Consequently, the disc for this code pair shows  $240/1418 = 0.168$  or 17%. Correspondingly, in Figure 6, the responses for the segments coded wall-technical were 37 out of a total of 194 word/word group occurrences analyzed, so  $37/194 = 0.190$  or 19%.

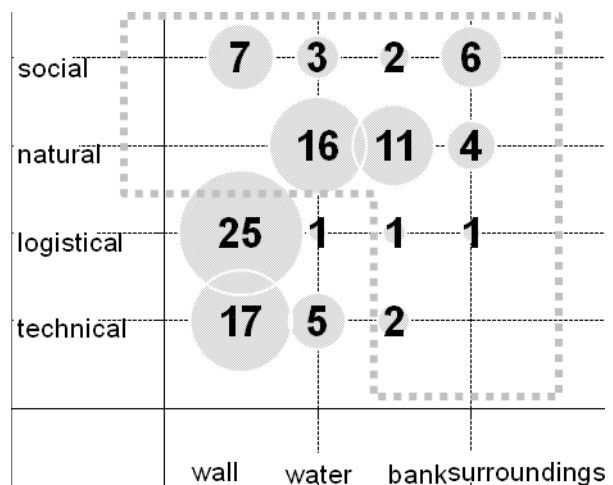


FIGURE 5  
DISTRIBUTION OF SEGMENTS ACROSS CODING SPACE, ORIGINAL STUDY.

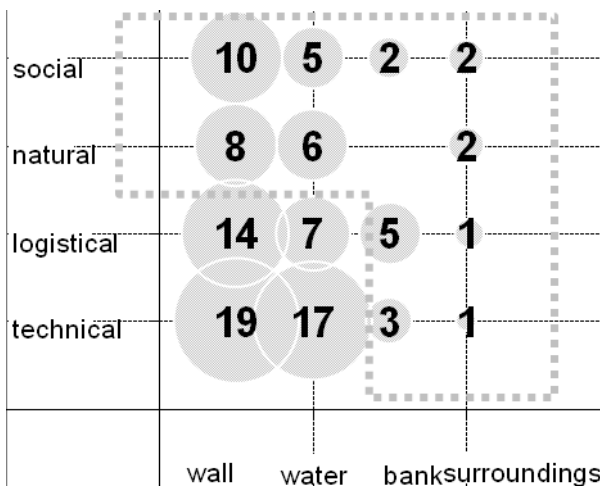


FIGURE 6

DISTRIBUTION OF WORD/WORD GROUP OCCURRENCES ACROSS CODING SPACES, HYBRID ANALYSIS WITH WORD GROUPS WALL+WATER+FLOOD+AFFECT/EFFECT.

As can be seen in the graphs, the data is dispersed in a similar manner with the exceptions (greater than 10%) of wall-logistical, water-natural, water-technical, and bank-natural. These discrepancies may be attributed to the limited number of comparison transcripts that were used (only 20% of the original study’s transcripts) and the limited number of segments analyzed (14%, assuming each high-frequency word/word group occurrence is considered a segment). As expected, given the word/word groups used to focus the coding in the hybrid analysis, there is some bias toward “wall” and “water,” likely accounting for larger percentages of detail-oriented words. However, even with this bias, a substantial minority of word/word group occurrences were coded as context-oriented.

When comparing the results of the original and hybrid methods with respect to the context-detail orientation, the results do not differ radically, especially considering the likely bias toward detail resulting from our choice of word/word groups. The hybrid method found 56.7% of the factors were detail-oriented, compared to 48.2% in the original study.

DISCUSSION

Based on the results of the comparison of the original qualitative analysis and the hybrid analysis, the hybrid method does not completely replicate the results, however, the general direction and trends are still captured using a

reduced number of segments (in this case, high frequency words) that are coded. Consequently, the hybrid analysis is better suited for larger data sets, as the overall aggregated numbers for context-detail were closer than the individual code proportions in the two-dimensional analysis (Figures 5 and 6), although the general trends were similar. This suggests that using this hybrid approach for focused, open-ended questions will capture the general ideas and trends versus using a traditional qualitative approach. The benefit to using this method is that the time for analysis is greatly reduced (provided the upfront work of developing the important word lists is completed) versus a traditional qualitative analysis. The tradeoff, however, is that you do not capture every thought and may lose novel ideas from a minority response. Consequently, this type of analysis is better for trend analysis and a coarse first-cut of the data. It also provides an intermediate means for analyzing large amounts of data with limited resources.

ACKNOWLEDGMENT

The Academic Pathways Study (APS) is supported by the National Science Foundation under Grant No. ESI-0227558 which funds the Center for the Advancement of Engineering Education (CAEE). Additional support for this work is from the National Academy of Engineering and the Center for the Advancement of Scholarship on Engineering Education. Special thanks to Cindy Atman, Deborah Kilgore, and Andrew Morozov for their work on the original study and valuable input, as well as Tina Loucks-Jaret for editing assistance.

REFERENCES

[1] Weber, R. P. (1990) Basic Content Analysis 2<sup>nd</sup> ed. Sage Publications. Series/Number 07-049 ISBN 0803938632

[2] Strauss, A. & Corbin, J. Basics of qualitative research – grounded theory procedures and techniques. (1991). Sage Publications. ISBN 080393251

[3] Sheppard, Sheri, Atman, C. J., Stevens, R., Fleming, L., Streveler, R. Adams, R. S., Barker, T. (2004). Studying the Engineering Experience: Design of a Longitudinal Study. In *Proceedings of the American Society for Engineering Education Annual Conference, Salt Lake City, Utah, June 20-23, 2004*.

[4] Kilgore, Deborah, Cynthia J. Atman, Ken Yasuhara, Theresa J. Barker, and Andrew Morozov. 2007. Considering Context: A Study of First-Year Engineering Students. *Journal of Engineering Education* 96(4): 321-334.

[5] TextStat 2.6 (Version date 10/10/2005) © Matthias Huning 2000/2005. mhuning@dedat.fu-berlin.edu <http://www.niederlandistaik.fu-berlinide/textstat/>.