Visualizing and Exploring Qualitative Research: Interview Transcripts

Aisha Kigongo
UC Berkeley – School of Information
akigongo@ischool.berkeley.edu

Vanessa McAfee
UC Berkeley – School of Information
vanessa@ischool.berkeley.edu

INTRODUCTION

Our goal was to create a visualization tool that could be used for data exploration by a group of qualitative researchers. We wanted to create web applications that provide a way to visualize the entire corpus of interviews and also drill down into the existing interview codes. We worked with researchers who were studying body image and weight stigma but our framework could be used to visualize other coded interview transcripts.

There are currently limited tools for creating and sharing interactive visualizations for results of qualitative research. Most of the existing Computer-assisted qualitative data analysis software (CAQDAS) have a high learning curve and high cost to use. Text visualizations are challenging and some common text visualizations like word trees fail to keep the data in context.

In an interview on the blog Qualitative Researcher, Dr. Stuart Henderson discussed the challenges with visualizing the results of qualitative research and emphasized that although there was no "Tableau for qualitative researchers"[3], there is still a need for visualizing qualitative data. Henderson states that the "the strength of visualizing qualitative data is it adds that wide-angle view and allows you to see the forest a little bit more." We think this is a wide open space with lots of opportunity for additional research and web based visualization tools like D3.js help with prototyping ideas in this space.

In this paper we will present existing CAQDAS tools and text visualization techniques, and present two applications for visualizing interviews. While working on this project we realized some of the unique challenges with working with human generated text (transcripts) and hope to continue working in this problem space.

RELATED WORK

Qualitative data consist of words and observations, not numbers. As with all data, analysis and interpretations are required to bring order and understanding. This requires creativity, discipline and a systematic approach [1]. After data collection, researchers read through text and find themes or issues that recur in the data. These become code (or categories) that may be ideas or concept [2]. This initial list of codes may change as the user works with the data since it is an iterative process and being able to visualize the coded data and compare gives the research flexibility to break sub categories into more categories or identify patterns within themes that can be merged together. The literature mentions that people perceive information in different ways and suggest that reading is not natural or innate but visually processing information is what people innately do [4]. So adding visuals to represent text data such as interviews adds a real strength of visualizing qualitative data as it adds a wide-angle view and allows the researcher to see the forest a little bit more. However less attention has been paid to how to display qualitative or text data visually [7]. This is probably because qualitative work focuses on explaining the “why” and “how” of complex phenomena, which are not easily portrayed with images.

New enterprise tools like Nvivo, ATLAS.ti, QDA Miner, Dedoose have come up that enable researchers to visualize qualitative data where they provide modeling or mapping capability where the user can link themes together [4].

CAQDAS Systems

Computer Assisted Qualitative Data Analysis Software (CAQDAS) systems like Nvivo, ATLAS.ti and QDA Miner are amongst the top enterprise packages available to qualitative researchers however they are also costly which make them prohibitive to many researchers, especially to students. A new tool, Dedoose has recently been released on the market to compete with the enterprise products.

Dedoose is a relatively new platform designed by researchers for researchers. It is an inexpensive web-based platform where a user imports documents, reads them, and creates codes or higher level conceptual themes and applies them to excerpts of the document. It has some visualization capabilities however it is not flexible in visualizing highly unstructured interview data from a multitude of angles and it is also still not freely available for researchers to use.

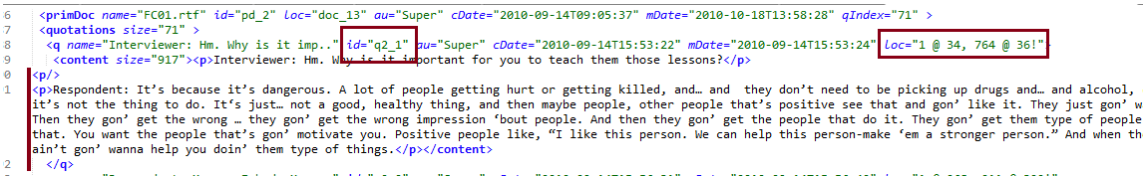

Figure 1. A sample quote from the XML file exported from CAQDAS tool ATLAS.ti.

Text Visualization Techniques

There is not a lot of research that has been done on visualizing interview data but a few important text visualization tools include Wordle [5], Wordseer [9] and IBM Many Eyes [6] which have made their way to the front and are frequently used in analyzing and visualizing qualitative data. Qualitative data can be visualized in many ways from individual words, sentences and context. Words are commonly visualized using word clouds, whereas word trees and mind maps are used to visualize sentences. Storyboarding is applied to contexts however all the visualizations rely on either frequency of words or removing stop words from sentences which tampers with the context of the interview [4].

The challenge in visualizing interview data is that any data visualization software and attention is centered on visually representing quantitative data [3]. Bar graphs and pie charts are great for conveying the meaning of large sets of numbers however there is a huge challenge in visualizing data mostly made up of words. Mechanisms for visually displaying qualitative findings have been underdeveloped and for those who do qualitative work, they have been limited to text organized by theme with a few quotes [3]. Shneiderman mentions that the purpose of visualization is insight not pictures [4] a statement that is agreeable with most of the literature. There has been less attention paid to how to display qualitative or textual data in visual form. There seems to be hesitation in how a researcher is able to maintain the integrity, tone, depth and essence of interview data in a visual representation. Some of the literature questions how a visual representation strengthens or weakens the qualitative data [3].

In our research we explore techniques for visualizing qualitative data (such as bar charts, brushing, mapping codes to quotes) and explore a variety of ways that researchers can portray their qualitative data.

METHODS

Data Cleaning and Processing

We received the data in the format of a collection of plain text files with interview transcripts, and XML file export from ATLAS.ti, and rich text files containing quotes. The corpus included the researcher's notes to themselves which we were not concerned with visualizing for the purpose of this project. The XML file contained quotes as shown in Figure 1. Each quote had an id number which also represented the corresponding interview number and also a location for where the quote was located in the plaintext interview transcript. Transcripts were created by many different people and the format for each varied.

After showing our first prototype to the researchers we learned that not all codes were of equal value and some were just personal memos to the researcher about the interview or the participant. We also had included codes such as "Transcription error" which was actually used to indicate mistakes and quotes that should not be used at all.

Processing Transcripts for Visualization

Our processing pipeline is shown in Figure 2. We processed the interview transcripts using a Python script which parsed the XML for codes and quote locations and also stored quotes in a MySQL database. The interview timeline required some text processing to determine who the "speaker" is (interviewer or respondent) and also to calculate the number of words in each response.

Interview Timeline View

We have developed a tool that allows the researcher to visualize the codes within one interview and also get a sense for the entire interview. Creating a word cloud of an interview pulls out the most frequently used words (often unigrams or bigrams) and is a technique for evaluating a corpus of text. The issue with word clouds is that often the context is lost. We wanted to include text in this visualization but also provide data about the interview timeline.

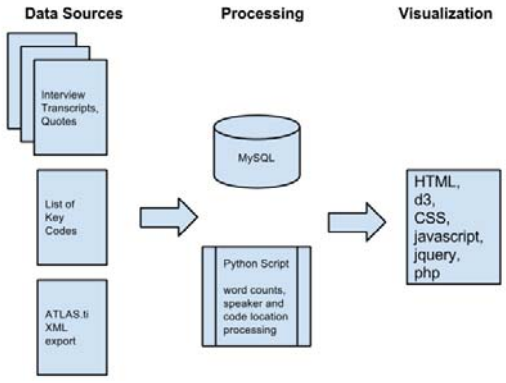

Figure 2. The process for preparing data for visualization.

Code Distribution across Interview Corpus

During data analysis, the interviewer reviews the data and demarcates segments within it. Each segment is a code usually a word or short phrase that suggests how the associated data segments inform the research objective. To visualize all the codes within the interviews, we wanted to create a word tree that would enable the researcher to navigate the data through different text. However that posed to be a challenge since the data was highly unstructured and consisted of a lot of slang words or incomplete or meaningless English exclamations that did not provide the same contextual meaning in a different interview. This made it difficult to work with and derive any insight within the data. Also the word tree was not a good visualization for codes that contained nearly the entire interview.

In the end for all the interviews we decided to explore visualizing text with text in a more structured way and additionally providing a quantitative view of the distribution of the codes between the interviews.

RESULTS

In this section we describe the results of our work, two applications one for viewing individual interviews and one for exploring the entire corpus. V.M worked on the interview timeline and A.K. developed the corpus view.

Interview Timeline View

The Interview Timeline loads with the timeline of the responses from the respondent displayed as a bar chart where each bar represents an individual response and the length of the bar indicates the number of words in that response. The codes that the researcher has used in the selected interview are loaded on the screen in a menu and when a user clicks a code, the associated response bars will be highlighted in green. The gray bar below the bar chart provides redundant position information but we believe it will help the user to perceive the location of the colored bars quickly. We found that sometimes the coded responses are actually very short resulting in small bars and they may not be easy to view. The uniform height bar chart below the response bar chart helps to draw the eye to that section. Note that we had difficulty drawing an axis on the chart but we should have one.

Figures 3 and 4 show the user interacting with the interview timeline.

Code Distribution across Interviews

The code distribution application allows researchers to explore the distribution of codes throughout the interview corpus. The interviewer can select a code to see the quotes associated with it and she is also able to view a quantitative distribution of the codes amongst the interviews at the bottom bar chart. For example, Interview FC02.rtf has the most quotes on Body Image. The interface is shown in Figure 5.

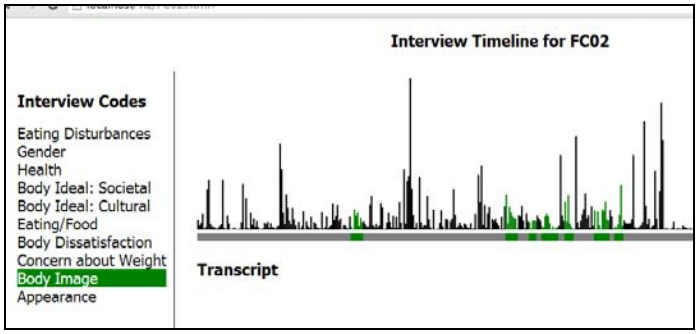

Figure 3. Clicking on the Body Image code highlights the seven sections of the interview timeline where the respondent discussed this topic. Bar length indicates the amount of words the respondent used in the answer. The gray bar below the graph helps the user to detect shorter bars faster.

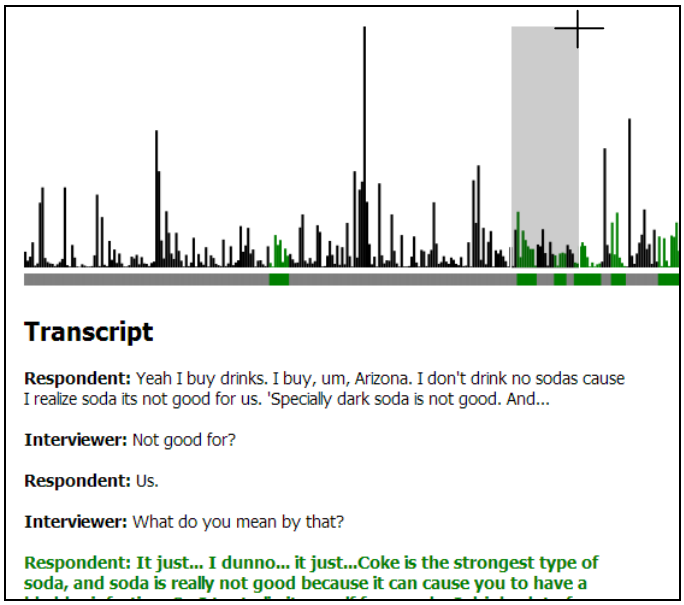

Figure 4. The user can use the brush tool view this section of the interview and view the context surrounding the coded quotes. The interview transcript text is updated dynamically below the timeline and the color indicates that it is a part of that code.

On the far right side of the application, the interviewer has access to all the interview files from which the quotes belong. By having a complete picture of the coded data, the interviewer is better equipped to prepare reports via a mix of summarizing the prevalence of codes, discussing similarities and differences in related codes across distinct contexts, or comparing the relationship between one or more codes.

DISCUSSION

Our audience has been receptive of the work so far done. They have expressed an interest in our approach to visualizing interview data especially the brushing technique within individual interviews. They also liked that the application provided an overall view of all the interviews and the flexibility to dive deeper and explore one interview.

Our audience feedback has been crucial in shaping the visualizations we used within the application. Comments like rethinking whether the data required a lot of interaction, suggestions to evaluate the word tree because it seemed to drop the context around the actual sentences and encouragement to look at tools such as WordSeer (http://wordseer.berkeley.edu/) [8] has helped us to tailor our tool and filter out inconsequential visualizations that did not add quality to it.

Overall the audience (students, a few qualitative researchers, and others) thought the application would be a useful tool to a lot of fields such as humanities that rely on lots of interview data.

FUTURE WORK

We have built web-based applications for visualizing coded interview transcripts on the web that allow a researcher to view individual interviews and also view the distribution of codes across interviews. We have also built the beginnings of a pipeline which can convert XML output from coding tools like ATLAS.ti into D3 visualizations. Future work includes making this a more streamlined system and giving the user the option to add new coding information. We would also like to be able to ingest a wide variety of types of interview transcripts. We have observed that interview transcript require a considerable amount of data wrangling because transcriptions are often done by multiple people and the data is semi-structured. We have also been requested to build a tool for visualizing the overlap between words contained in codes. We could also look into topic modeling algorithms and reveal suggestions for responses that could be used for quotes directly in the interview timeline interface.

ACKNOWLEDGMENTS

We thank researchers at the Center for Critical Public Health in Berkeley for allowing us access to real data to work with. We hope to continue working with this group to develop other visualization tools to increase collaboration within the organization and externally.

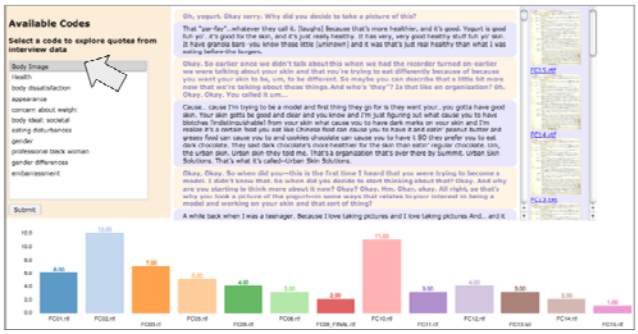

Figure 5. Interface for the code distribution application. Codes are show in the left panel and the middle section displays the quotes corresponding to the selected code. The right panel provides a way for the researcher to view the interview document. The bar graph displays the distribution of the selected code across the interview corpus.

REFERENCES

1. Taylor-Powell Ellen, Renner Marcus, Analyzing Qualitative Data, 2003, University of Wisconsin-Extension, Madison, Wisconsin, Program Development & Evaluation.
2. Ratcliff, Donald. 2002. Qualitative Research. Part Five. Data Analysis. http://www.don.rat-cliff.net/qual/expq5.html.
3. Elliott, Susan. 2012. Qualitative Data Visualization: An Interview with Dr. Stuart Henderson. Qualitative Research http://www.qualitative-researcher.com/qualitative-analysis/qualitative-data-visualization-an-interview-with-dr-stuart-henderson/.
4. Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman, eds. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
5. Wordle, www.wordle.net
6. Many Eyes, http://www-958.ibm.com/software/data/cognos/manyeyes/
7. Henderson, S. and Segal, E. H. (2013), Visualizing Qualitative Data in Evaluation Research. New Directions for Evaluation, 2013: 53-71. doi: 10.1002/ev.20067.
8. Muralidharan, Aditi, Marti A. Hearst, and Christopher Fan. "WordSeer: a knowledge synthesis environment for textual data." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.
9. CS 294 Students comments, http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/index.php/FP-AishaKigongo-VanessaMcAfee