Introduction

Much of qualitative research is concerned with meaning-making processes or schemas (i.e., cognitive meanings). Specifically, many researchers are interested in cognitive perceptions and how respondents create and use meaning of the world to guide behavior. While qualitative research is currently the best method to mine deep meanings and cognitive schemas, it is time consuming and subject to high variability in researcher skill in recognizing and pursuing key schemas. Comparisons and use are thus tedious and potentially capricious.

I seek to contribute to our understanding of how schemas are organized and drive behavior, by bridging network analysis and social theory with techniques that mine language for the consistent patterns representative of schema. With network text analysis (NTA) techniques, I can map consistent language use within and between rich textual transcripts, and link these back to families of schema. This approach should provide insight to the measurement of schemas that are contradictory and thick, but are also less time intensive than qualitative research demands and these methods will show key themes that researchers may have missed. Using this method we can link words to their local contexts, to get not just volume but co-occurrence, which can give us structure of meanings.

The underlying assumption of NTA is that, as Carley (1994) asserts, words have no meaning except in relation to other words. NTA gives a roadmap of what words are connected to generate meanings. In this abstract, I examine which words are used in the same paragraph as ties. This links words to their local contexts, to get co-occurrence of words, which gives us structure of meaning (or patterns of word co-occurrence). This can then be combined with social network techniques to describe patterns in data, map out schemas, and to test hypotheses.

I believe this method, adapted from social network techniques, has a great deal of promise. At the very least, network text analysis can simplify qualitative research and help researchers figure out where to look to make the process less time consuming. However, my hope is that the methods described above will become a tool to examine schemas that bridges the gap between quantitative and qualitative analysis. This method gives researchers the tools to use qualitative data to map out schemas, quantitatively describe and test hypotheses with text data, and to combine the thick description of qualitative analysis with the replicability and formal hypothesis testing that quantitative research offers. NTA gives contradictory and thick descriptions with numerical representation while being much less time intensive and allows researchers the opportunity to use larger resources of text than has ever been possible. Additionally, with NTA new research questions and themes emerge which broaden research both methodologically and theoretically. I hope these efforts can bridge the qualitative/quantitative divide while enlarging and transforming the scope of possible answerable questions for all social scientists.

In this abstract, I will show how these methods can be useful when examining qualitative transcript data. I use an example using data from Becoming Parents and Partners study (BPP), a qualitative in-depth interview study designed to examine the meanings of childbearing for low-income Blacks that had not yet experienced these transitions, to examine how NTA can aid in and add to qualitative methodology to understand cognitive meanings of respondents.

Data

The Becoming Parents and Partners (BPP) study was designed to examine the views of unmarried, childless young adults toward marriage and childbearing. The motivation behind the BPP was to explore the attitudes and expectations of marriage and childbearing among individuals who had not yet made either of those transitions. The interviews took place in a mid-size southern city, and were fielded in the summers of 2009 and 2010. The respondents were non-Hispanic African American, between the ages of 18 and 22, unmarried, without children, and not currently enrolled in a four-year college. The interviews took between 45 and 90 minutes, and interviews were taped and transcribed. Respondents were asked about the meanings and norms surrounding marriage, cohabitation, and childbearing; their aspirations and expectations for marriage and childbearing; the level of expected self-fulfillment from being married or having a child; their families' expectations and experiences with marriage and childbearing; their peers' expectations and experiences with marriage and childbearing; and current relationships.

Method
The goal of NTA is to examine the meaning context of language by examining how words are connected to other words. Based on the premise that words have no meaning except in relation to other words (Carley, 1994), NTA maps out how words are connected, not only in volume, but more importantly, in co-occurrence. Once co-occurrences have been established, social network techniques can be used to describe patterns in the data, identify schemas, and test hypotheses.

NTA is a several step procedure including preprocessing, calculating tie values, clustering, calculating betweenness centrality, and visualizing the text networks. These steps are described in more detail.

Preprocessing prepares the textual data for analysis and allows for calculation of co-occurrence between words (known as "ties"). Preprocessing involves taking the text of the interview, deleting the interviewer's questions and responses, and retaining only the text of the interviewee. Each interview response constituted one paragraph, even if the response grammatically constituted more than one paragraph. Next, the text was cleaned and codified, which consisted of automatically replacing synonyms and stemming words to their roots (e.g., "want", "wanting", and "wanted" are all stemmed to "want"), and removing most non-informative words such as pronouns, auxiliary verbs, and interjections. All identifying information such as names and places were also removed. Because including terms only used by few respondents would make ties look stronger than they actually are and does not show shared group meanings, terms were deleted if they were used by less than 15% of the respondents (10 people). Additionally, I deleted terms that occurred in 90% or more of the paragraphs, because very common terms such as 'um' and 'like' tend to have little meaning. Slightly raising or lowering these percentages does not substantively change our results. In total, I had 563-stemmed terms in 17,562 paragraphs.

After preprocessing, I calculated tie values in the following manner. SAS's Text Miner package was used to create a sparse term-document matrix. In a sparse-term document matrix, each term in each paragraph constitutes its own entry (term$_{ip}$ for term *i* in paragraph *p*, term$_{jq}$ for term *i* in paragraph *q*). I then counted the total number of paragraphs in which term$_i$ occurred, and the total number of paragraphs in which term$_i$ and term$_j$ co-occurred. These totals are divided by the total number of paragraphs in which *i* occurs, resulting in a percentage of how often *i* and *j* co-occur in paragraphs that mention term$_i$.

A tie is computed based on equation 1.

Eq. 1.

$$co\text{-}\ occurrence = \sum_{p=1}^{p} term_{ij} \Big/ \sum_{p=1}^{p} term_{i}$$

As shown in Equation 1, the strength of ties is the total number of paragraphs term$_i$ and term$_j$ have co-occurred divided by the total number of paragraphs term$_i$ has appeared in. This creates a directed (i.e., asymmetric) network of connections where two words can have asymmetric ties, as shown in Equations 2 and 3.

Eq. 2.

$$\sum_{p=1}^{p} marriage \,\&\, child \Big/ \sum_{p=1}^{p} marriage = .33$$

Eq. 3.

$$\left. \sum_{p=1}^{p} marriage \,\&\, child \middle/ \sum_{p=1}^{p} child \right. = .23$$

In equation 2, the term child appears in 33% of paragraphs that contain the term marriage; in equation 3, marriage occurs in 23% of the paragraphs that contain the term child. I calculated the asymmetric ties for all pairs of words, resulting in a network of connections between terms.

In the next step I cluster terms to analyze how words are grouped together. I clustered terms by using community detection algorithms which have traditionally been used in the social network analysis literature (for methods reviews, see Porter, Onnela and Mucha 2009; Moody 2001). Community detection algorithms involve dividing terms into mutually exclusive groups to maximize the number of ties *within* each group and minimize *between* group ties.

The clusters show what concepts are connected to other similar concepts. This measures what clumps of knowledge exist in individuals' heads. Labels for each cluster are decided based upon the concepts in that cluster. I guide my labeling by identifying the most central concepts within each cluster. Because clustering algorithms cannot be calculated for directed networks, I symmetrized the matrix to the maximum tie value; in our previous example I showed asymmetric ties (child→marriage at with a tie strength of .33 and marriage→child at a .23 level), to symmetrize the network I chose to use the maximum value of the tie, or child←→marriage equals .33.

After clustering, the next step was to calculate the centrality of the cluster, which demonstrates which words are most important in connecting the text networks. I calculated betweenness centrality for each term. Betweenness centrality is a numerical measure, which shows how much a specific term is between (or connects) all other pairs of terms in the network. This measure shows how important words are to connect the text networks; in other words, betweenness centrality shows how much the text network would break apart if that specific term was deleted from the network.

Finally, I visualize the text networks using network software (specifically, Pajek).-In these visualizations, terms are represented by circles and the strength of ties is represented by a line connecting terms that is darker for higher tie values. The colors of the circles which represent terms shows which cluster each term falls within and the size of circle expresses the betweenness centrality of that specific term. Terms that have higher betweenness centrality, or are important to connect the term network, are larger.

In the following, I show a figure that represents the child-term network. These show all the words connected to child or marriage at a level of .3 or above (only words that co-occurred with either child or marriage in 30% or more paragraphs).

Results

Figure 1 shows the child-term network partitioned into clusters. The colors indicate which cluster a group of terms falls into. Child is not shown because all terms are connected to child. There are 273 terms that are found with child in 30% or more of paragraphs. Of those 273 terms, 113 are connected to at least two other terms. Here I only show words that are at least connected to two other words.

Figure 1. Child-term Network

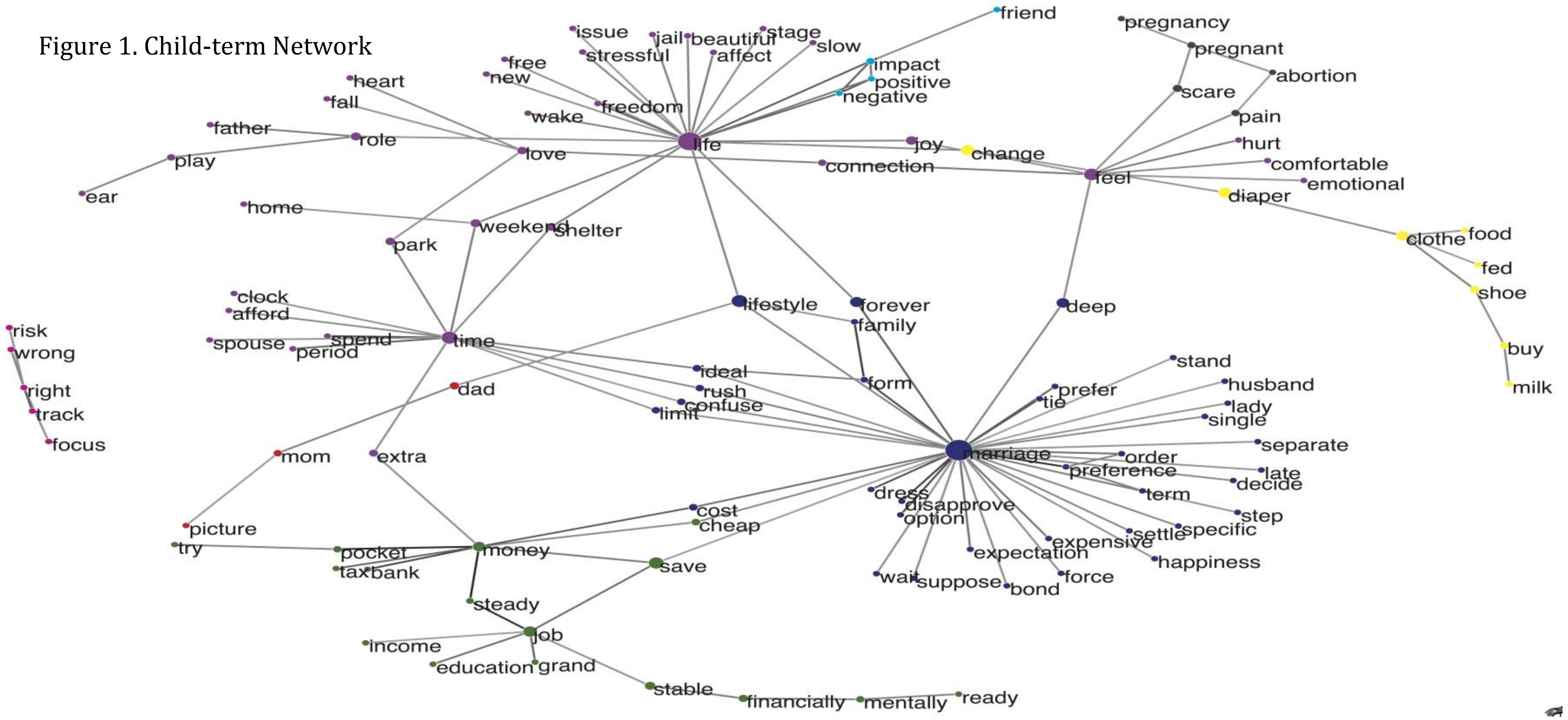

Figure 1 shows the schemas that are connected to child. We see that marriage is connected to childbearing because it is represented by a large cluster within the child-term network. This cluster (illustrated in blue) describes marriage's connection to childbearing. Also, financial considerations are important before having a child for those who have not yet had a child (green) and respondent's report that emotions and the time required for childrearing is important in childbearing decisions (purple). Also we can see that respondents describe the things needed once one has a child (yellow).

Discussion

This shows the schemas, or cognitive meanings that surround childbearing, that this method can be a useful method to show meanings of concepts, connections between concepts, and can empirically test for differences between groups in how concepts are defined and connected.