The Application of Concept Mapping to Text Analysis: Examples and Validity Issues

Kristin Behfar

I want to talk about concept mapping. Concept mapping is an alternative to some of the methods presented so far. I like to talk about the range of what it can accomplish. There are many different types of methods that go by the name concept mapping, many different variants. One is primarily visual/drawn – it comes from political science, Axlerod, one of the earliest demonstrations of cognitive maps of political elites. Also in education they deal with connectors – if/then and relationships along a visual way. The more statistical variant, which is semantic mapping that we have heard a lot about. What I am talking about today is sort of a blend of the two. It consists of 5 steps. It includes participant generated units and then a combination of human judgment with computer statistical analysis. It is also kind of a hybrid of word vs. code type approach. We have a unit of analysis, a statement that is broken down from a larger open-ended survey response and then we use that as a unit of analysis and based on decisions that you make in steps in the analysis uniforms it to a great degree or not.

In terms of application to open-ended responses I got thinking about this because many of the questions I ask are deductive type questions. Especially when it is dominant topology or an existing coding scheme in the field that is not necessarily working you need something that allows data to emerge rather than to force into pre-existing categories. The characteristics of the text that I usually have to deal are similar to open-ended responses. This makes semantic mapping a little bit of an overkill because there are very simple relationships embodied in the units of analysis. Then the forced categories (content analysis approach) don't really get to the research question that I am asking. So this is why the hybrid approach to this method is particularly well suited to open-ended questions.

(Pointing to reference on slide) Jackson is my maiden name. That is me. In addition, Bill Trochim was the brains behind the methodology. He was my dissertation supervisor. If you have questions about the nitty gritty of the methodology he would be a great person to talk to.

I want to step you through the 5 steps and then I will talk a little bit about reliability and validity and then kind of show 3 examples of extensions of the base analysis – how to answer research questions a little better.

The first step is unitizing. You take the respondents examples and break them into units. The key here in terms of mapping is that every unit can only contain one idea. It is based on sort methodology so when you have more than one idea on a card it is hard to know whether you put it in pile a or pile b. So every unit has to contain one statement. That is a limit of this methodology. You can't, for example, map if/then statements. You can't do negatives and positives on this type of map. The way that I have handled that in the past is if there are negatives and positives I separate them into different analyses. The other thing is that this is one unit – here is an example of the most important issue – just from one person – abortion, immigrants coming in, military and gay rights. That is 4 from one person. If you have 4,000 people that is a massive amount of units. The problem is that you put them each on a card and have the sorters sort them usually that way exceeds the sorter burden. In my experience is has been 100-300 statements that people are capable of sorting in one sitting. More than that and they get cognitive overloaded. The way that I have dealt with that in the past is – here is a question about MBA students who were asked what kind of conflict they had on their team. Some of them said "none," some said "none yet," and some said "none to speak of." They all went into the none cluster. The problem is that if you want to retain the nuances. What I did was before the mapping analysis I did a key word context analysis to come up with the word none. Then I identified from the nuances using my own judgment between those three categories and I created placeholders in the database if you have the word "none" – for example, if there are 50 associated statements and I am going to take those units out a sort deck-- but when they go into the cluster then I put them back in the lists or the tables or the database that I create. The problem with that is that if you take out too many statements, participants might be sorting them with other concepts, or create different types of categories. For example, if there are 500 statements about abortion and you only leave one of them in, and only leave one in about the death penalty—someone might put them together in a "right to life" cluster rather than separating them.

What you really want to know is that there is an abortion cluster and there is a death penalty cluster. The rule I use there is that I leave 20% of the total in. I have never been asked to justify the 20%. The way that I think about it comes from the literature on affirmative action. For example, they find that 1 in 5 or 20% is the point of optimal distinctiveness. The other rule of thumb is to use is a minimum – for example, if in one category I take only leave10 statements– I always leave the same minimum of all the rest of the categories. That is one way to get around the scaling-up problem (to avoid cognitive overload in sorters).

Here is an example response that I broke into 2 units. The key here is that if you leave more than one idea in a unit it is almost impossible to sort into a discrete pile – and in terms of multidimensional scaling, you get a junk cluster in the middle. Unitizing is very, very critical. The way I do it, usually it is a very time consuming process. The whole analysis would probably take a couple of hours but the unitizing takes a couple of days depending on how many statements you have. I usually have 2 researchers make the unitizing decisions and then access inter-coder reliability. Or we have also very successfully used a group of participants who generated statements – 3 or 4 of them get together and make collective decisions about what constitutes units. I have done that on interview transcripts and on open-ended question surveys. It depends on how much of the participant judgment you want to retain and what you think your reviewers will think about the validity.

The second step is to sort. Here is a hypothetical sort of 10 cards. The sorters that I choose are usually the original participants. Sometimes there are confidentiality issues so I can't use original participants because for some reason they can identify – in the MBA section – who said what about who. We remove all the names and company references – I usually use the original participants, a minimum of 10, but I like to get about 24. If you can't use original participants we have what we call proxies. The criteria we use in identifying proxies are how much of the original experience do they share in terms of the validity and the interpretation. If they use words that they don't understand the context of the words, the meaning, then it is very difficult to run a sort in the same way that would reflect the original structure from the participants. The second is how much knowledge of theory they have. When I was talking about what conflicts they have on their teams, I had the challenge of after a certain point they get an in-class module on managing conflict. In which case they would reflect the dominant topologies that we teach and the structure wouldn't necessarily emerge from their perspective. I had to make sure I used people that hadn't had that theoretical background. That is where my research interest was. The third was how much existing theoretical frameworks can guide your interpretation of what they are doing? So for example, in team management we know typically there is a task side and a social side. You manage the task and you manage the people where you work. You use those taxonomies to understand how much the sorters are completely blind or how much they do know what they are doing. You can identify people who aren't following instructions or aren't taking the sort seriously. Here is a hypothetical sort – the instructions are to create as many piles as you think conceptually makes sense. It is not queue sort with forced distribution. The only rule that they have is that they can't create a junk pile. If they don't' think a statement belongs in a pile they leave it in its own pile and give it a name. When we aggregate by adding together all the sorter's initial matrices—the result is a point map, where the collective (or aggregate) judgment decides where that point goes on the map. The sorters have labeled each of their piles. For this one – 6, 2, 1 & 9 go together so we create a binary score matrix for each of the sorters. In the database this matrix shows what the piles were and each of their piles is associated with a label. Then we aggregate just by adding all the matrixes together. You get an aggregate and you do an MDS on that, an aggregated matrix and you get an aggregated picture of your relationships between the statements and the units.

I'll show you a few examples of what the statements were so you can see how they fit together. [Pointing to slide] "Conflict on the best way of solving a problem"; "we didn't know how to organize or approach the work;" "different views of the case;" this is mostly task related; "one person always having to be right, sometimes oppressive;" "tremendous personality clashes;" "no open fighting just bottled-up resentment occasional sarcasm;" so there is overt kind of implicit conflict; "some felt lack of ownership;" "conflict regarding team members showing up late or not showing up at all;" "time spent together was too long." Task issues, personal issues, logistical types of issues. So the next step here is to decide what to do cluster analysis on these coordinates. We have to decide what is the final cluster solution which is another step in the analysis – in choosing who will make this decision, you have to decide how much theory will inform us vs. how much a naive participant will inform us.

Here are the results of this one. In this case I used two of the original participants to decide what is the cluster solution. So, the "best" cluster solution is a matter of human judgment. I chose the original participants because my brain is already saturated by the dominant taxonomy in the literature. In instructing the decision makers, what I basically do is print out the dendogram and all the related text for each of the points associated with each cluster solution. They work together to make the choice of a solution– when does it make sense that the clusters merge and when does it make sense that they stay distinct? They chose this number of clusters. I have also done it where I use different demographic groups to see if they come up with different types of clusters. For example, I have asked my professional colleagues do the clustering so we could compare their clustering solutions to participants. Then we then conduct interviews as to what those differences are and their interpretations-or you can assess correlations between 'expert' and 'novice' matrices.

The next step is to label the clusters. There are a couple of ways to do the labeling. One way is to have the humans choose it. I typically have the decision makers of the cluster analysis choose them. They are very familiar with the data. The other way, the more statistical way, is to do a centroid analysis where you come up with the average X & Y values for each of the piles that each participant creates that is associated with their label. You come up with a centroid for each cluster. Then you come up with a top 10 list using the equivalent distance for each person's pile – which one is the closest to the center of the cluster. You can generate into a "top 10 list" of names for each cluster. Usually I give them a top 10 list and the participants use that to inform their judgment. Usually they want to put some kind of variant on it.

Here are the results of this. In assessing reliability there are a couple of things – I use Klaus Krippendorff's categories. I have actually never been asked any of these questions by a reviewer, ever. But in terms of stability there are a couple of ways – the same coder being consistent. You can ask sorters to do sorts at two different times and then correlate the matrixes. Reproducibility is a little more difficult. One way is to correlate each individual matrix with the aggregate, like an item total assessment. This is really dependent on my choice of sorters. Often times a research question you want to know is this reproducible or not?

You can compare experts vs. novices and individuals vs. the aggregate. Accuracy is very hard. There is really no solution in this case. I like the other solutions. This is where I think the method can shine. Some units are harder to code than others – that goes away because the participants are doing it. If they are hard they are left in the middle and pulled toward the aggregate. When they do show up in the middle of the map and usually in my experience it is because someone made a bad unitizing decision where it is hard to interpret units. In which case you take a second stab at unitizing it or try to explain it. In some cases it is theoretically interesting. I will show you examples of that. In other cases where it does truly reflect to researcher error but that has to be justified. Some categories are harder to understand than others – this kind of goes away also because the participants are creating the categories. Subsets of categories can be confused with larger categories – that also does go away in the aggregate picture. You lose some fine tuning but you get the aggregate picture. I will talk a little bit more about that later. Individual coders may be careless, inconsistent or interdependent. The interdependent goes away. There is no conflict resolution over where this unit belongs. But coders can be careless for example they don't follow instructions. One coder I gave him a list of conflicts with an MBA team and he divided the statements into two piles of conflict: 1) that he personally has experienced versus 2) those he hasn't experienced-- rather than to create categories of conflict. People can be careless or not follow/misunderstand instructions. That is where the matrix correlations from individual to aggregate can really make this apparent. Another example: one of the sorters used a taxonomy he learned about in undergrad that describes different stages of group development: forming, storming, norming, performing. It is a taxonomy sometimes used to describe stages group development. It didn't necessarily rely on conflict at all so we threw that one out as well.

In terms of validity, this really depends on the choices that you make. Sampling is a big issue here. I have done a lot of research with concept mapping on team level analysis in which case it is very difficult. You either have all the team members and find a way to access team level agreement, or use 'informed experts' (members who represent their team). I'll show you a way I have assessed within and between team agreement. But one problem with 'informed experts' is if you have one person reporting on what happens, for example, I did a study on negotiation teams, we had one expert informant. You have to make it very clear why the expert informant would represent the team. In this case we sampled from executive education so they knew the difference between what was within the negotiation team issue vs. what was across the team issue (an important internal validity concern). We actually wanted them to have some familiarity with theory and practice.

For retaining context it is semantical validity issue. The unitizing is key; this is where the reviewers go after me the most. Assessing reliability is very important. The statements that show up in the middle of the map, that is where I tend to focus when I have validity discussions. Is it really a unitizing error or is there some reason why – [pointing to slide] there are some types of conflict that are personal but they are personal because of cultural differences – those are the ones that show up in the middle of the map because there are two reasons why there is conflict in that statement but they are hard to separate. That is very difficult.

The choice of sorters is important for external validity: who are you trying to generalize to? The structure that emerges has to represent the population that you are trying to generalize to. In my case the studies that I do I tend to use census, all the MBAs that are on this type of team; all the people that do the negotiations.

Internal validity is almost like a reliability issue in this type analysis.

External reliability is critical to the interpretation. This has been tricky both as a researcher, but also when presenting results to participants. As a researcher you rely on the proximal distance and the proximity of clusters to – does the proximity of how the participants sorted and viewed different clusters and categories of conflict jive with the categories in conflict in literature and other taxonomies. We understand how to interpret multidimensional scaling results. But, audiences and participants do not necessarily interpret the maps well. For example, in terms of participants it is very interesting; a colleague of mine was involved in a search for a new police chief in the city. Part of the problem was that there was a lot of racial tension in the community between the police chief and the police force and the community. They solicited from the entire community (using newspaper ads and suggestion boxes around town) what are the top 5 qualities that you want to see in a police chief. They did concept mapping. They went to the community in an organizing type meeting. They put up the concept map and immediately people started booing. The "racial sensitivity" cluster was a very small cluster down at the bottom. The audience figured that since it was small and at the bottom it wasn't very important. They didn't understand that it is the spatial relationships that matter. The size and the shape of the cluster really is not that meaningful. Smaller clusters tend to be tighter in terms of conceptual clarity but the bigger clusters don't mean more important. That is a drawback that I couldn't have foreseen about this type of methodology. You have to be very clear about how to interpret the maps—in some cases I would recommend to not show them—just use a list rather than the visual representation if you think there is any risk.

Now I am going to show you three examples that demonstrate the range of the methodology. You have the map which is basically a categorization of participant judgment or structure in data. I am going to give you three examples on how to expand that map. I just published a paper on these categories. Here is the same map I showed you about the types of conflict that exist. I wanted to compare this to the expert judgment, the taxonomies that exist, dominating the literature. This is where it is particularly interesting. There are three types of conflict that we talked about – task, relationship and process. The problem is that task conflict and process conflict never separate with factor analysis, so people just leave out the construct process conflict. However, process conflict is a critical foundation of theory building in the team's literature. Look at what I found – these are very close [pointing to the task and methodology/approach clusters] – I had experts, academics do Likert scale ratings on each of the statements as to how much they were related to task, how much they were related to relationship and how much they were related to process conflict. Then I did a T-test to see how if there were significant differences among the ratings for each item – aggregate item – in each of these clusters. What I found is most significant differences between task and process conflict in this cluster or this cluster [pointing to the timing/scheduling and workload/contribution clusters]. The participants judge these to be very close together. The experts could also not tell them apart. So when you look at the current scale that measures these things, they are measuring this [pointing to the work method/approach cluster] but ignoring this [pointing to the timing/scheduling and workload/contribution clusters] and really these two are the unique aspects that cause this conflict. This comparison of participant vs. expert review was very useful.

We also used these clusters to generate scale items and to guide open-ended interviews based on these clusters.

Here is another comparison [next slide]. We wanted to see what are the challenges that multicultural teams have in comparison to the challenges that same culture teams have. We asked participants from two samples (same and different culture teams) to generate different statements and had them do different mapping. Based on theory, we discuss these two tables that represent participant view about what is the same and what is unique based on culture.

In this example [next slide] I wanted to tie three data sources from a survey together: 1) they type of conflict experienced, 2) how they resolved that conflict, with 3) team outcomes of performance and satisfaction. This is a picture of the map from how they resolved conflict. One of the clusters was that "we avoid it". This is a graph plotting team Likert rating scales for satisfaction against the performance or grade what they got on a midterm—for the statements in the "avoid it" cluster. I looked at the differences in the quadrants. I think looked closer a the qualitative differences in the quadrants. What I found was that the teams that were low performance/low satisfaction, they were defining avoid as "sweeping it under the rug/divide and conquer/minimize our misery." The high performance/high satisfaction people were defining avoid as "working hard to prevent." The low satisfaction/high grade which was the majority of the people in the class, defined avoid as "preventing escalation." The taxonomy, the dominate taxonomy, of conflict and management has avoidance as one of the strategies but when it is aggregated to the team level it doesn't capture the nuances of what the word avoid means depending on the experience and the tradeoffs they make on conflict resolution.

Because I was trying to capture team-level strategies, I had to assess within-team agreement. A challenge with qualitative data. The way I did this was to use the fact that we have three criteria for evaluating team viability one is performance, one is satisfaction and one is that the team's process has to enhance its ability to work together in the future. So, on this graph, one quantitative dimension is satisfaction and the other is performance. What I am plotting is process—or the results of the concept mapping analysis. To assess within team agreement, all members had to report experience the same conflict types and associate them with the same strategies. To assess inter-team agreement about which strategies were applied to which type of conflict, at least 1/3 (to reflect that there are three types of conflict) had to report applying the same strategy to the same type of conflict.

The last example I wanted to show you was how you predict behaviors and attitudes from the fact that a participant has a statement that belongs in a particular cluster. This is a study I did on negotiating teams. There is broad literature on teams and broad literature on negotiation. But there is very little research on how challenges within these teams impact their across-the-table strategy. We did two maps on what were the challenges they have and how did they manage them; not across the table, only on the same side of the table. When we unitized the interviews and we had to throw out all the across the table utterances. We unitized by going through the interview transcripts. Three researchers used highlighters to say this is a challenge, and marked the associated solution. In our database we had the unit, the challenge, the associated management strategy, their home industry, the number of people on the team – all those things we had to control for. Then we did a thing called pattern coding. You take smaller number of categories and create meta-categories. Based on existing theory, we came up with "global" categories represented by spatial regions on the maps. Each of these clusters [pointing to map] belongs in a meta category. The problem with doing that is – and the problem with interview data – if you get a particularly verbose person you might generate many, many units and this can distort the distribution of meaning in a cluster. What we ended up doing was constraining participants to one statement in each of these clusters. For example, if a person had 3 statements in one cluster, we only allowed 1 of those statements to be in the analysis. However, if a person had one statement in each of 3 clusters in a meta-category, we allowed it. A person that clustered a statement in all three of these, we allowed them but if they

Then we did a Chi square analysis to see if you are in this meta-category what is the likelihood that you will be in another midi-category? Then we looked for interdependencies because all teams have multiple challenges. We looked for interactions to see if you have interpersonal differences that affect your ability to find a resolution to task differences. What we found, contrary to what the negotiation literature would say, is that if you have task differences (that you are not able to formulate coherent same side strategy) – they actually did better across the table and had more positive receptions of the process, because by engaging in that strategy discussion the range of issues on the radar are expanded so they tend to address the interpersonal issues in a much more constructive way. The way that they did it based on the matches that we made across the maps was that they rehearsed. If so and so said this… That way they were able to maintain team discipline and prevent the other team from dividing and conquering them which is a common strategy between groups.

These are the practical issues I want to cover because I am almost out of time—they represent the advantages, disadvantages, and range of the methodology. It is a good fit for the nature of open-ended responses; especially with the organization surveys that I do that tends to be kind of a free listing context type of test. They don't rely on preconceived coding schemes. The research questions I was asking were primarily exploratory or inductive; I wanted to see what was wrong with the dominate taxonomies and why are they not empirically supported? The method preserves the context of the concept in the unit of analysis. As I showed the nuances of the word can be quite different. It is the same word. There are also considerable time savings – the unitizing takes the longest, sorting typically takes 20-60 minutes, it takes decision makers no more than 2 hours to choose a final cluster solution and labels. It is nice compared to, I was once involved in a content analysis project that took us three weeks of fulltime work. It is a considerable time savings.

For the challenges – you have to have access to the units of analysis. You just have to. There is no way to do it without them. There is a sorter burden. If you get beyond 300 units it can be difficult. Complex text, I tried to do this with restaurant hiring data. "I would hire this person if he had these three qualities but not this one." You can't map that with this. You need a more sophisticated mapping analysis for if/then relationships. Dense text is particularly difficult to unitize. In the interview transcripts that we unitize you often get a paragraph so the sorting endeavor becomes a different kind of sort. It is not just concepts, but it is putting practices into categories rather than just concepts. I have found that you have to have more sorters when the units are more complex. The other thing with dense text that reviewers have had trouble with – they often look at the stress value of the multidimensional scaling as a reliability estimate and it is really not because I am constraining it to two dimensions. It isn't a good indication on how much agreement there is among the sorters. Thank you.