SOME PROMISING RESULTS OF COMMUNICATION-BASED AUTOMATIC MEASURES OF TEAM COGNITION

Preston A. Kiekel, Nancy J. Cooke, Peter W. Foltz, Jamie Gorman, Melanie Martin
New Mexico State University
Las Cruces, NM

Some have argued that the most appropriate measure of team cognition is a holistic measure directed at the entire team. In particular, communication data are useful for measuring team cognition because of the holistic nature of the data, and because of the connection between communication and declarative cognition. In order to circumvent the logistic difficulties of communication data, the present paper proposes several relatively automatic methods of analysis. Four data types are identified, with low-level physical data vs. content data being one dimension, and sequential vs. static data being the other. Methods addressing all four of these data types are proposed, with the exception of static physical data. Latent Semantic Analysis is an automatic method used to assess content, either statically or sequentially. PRONET is useful to address either physical or content-based sequential data, and we propose CHUMS to address sequential physical data. The usefulness of each method to predict team performance data is assessed.

Team cognition is more than the sum of the cognition of the individual team members. Instead, team cognition emerges from the interplay of the individual cognition of each team member and team process behaviors. For the purposes of this paper, any small group of people collaborating on a task constitutes a team. The typical strategy for measuring team cognition is to define it as an aggregate of individual cognition, such as the average. In contrast, if viewed as an emergent property of the team as a whole, team cognition should best be measured holistically whenever possible (Cooke, Salas, Cannon-Bowers, & Stout, 2000). Communication data constitute an inherently holistic measure of the team. Further, communication serves similar functions in teams that cognitive processes serve in individuals. The output is also analogous to individual verbal reports. Thus, an analysis of team communication provides a window through which to view team cognition (Kiekel, Cooke, Foltz, & Shope, 2001). But many communication methods are time consuming to employ, and it is often difficult to achieve adequate inter-rater reliability. This paper outlines several relatively automatic methods for analyzing communication data, and considers their usefulness.

It is useful to characterize communication measures along two dimensions: "physical" data vs. "content" data, and "static" vs. "sequential" analyses. Physical measures are relatively low-level measures such as duration of speech. Content measures account for what is actually being said. On the other dimension, sequential analyses are those that account for the fact that utterances are only made in the context of an ongoing stream of team interaction. Static measures consider the team only at one point in time, or as an aggregate of the team's communication over some duration.

Toward the objective of automation and to collect physical data, we developed software that records quantity of verbal communication as an NxK^2 communication log (CommLog) matrix of dichotomous values. K is the number of team members, and N is the number of time intervals (e.g., seconds) across which the communication spans. All possible pairs of the K speakers account for the K^2 columns. At each time interval, a measure is automatically taken of which team members are talking, and to whom. This creates the N rows in the CommLog matrix. The result enables rapid analyses of sequential flow.

Content data were taken from transcripts of each team as they interacted in a simulated military environment across several missions.

METHOD

This paper shall focus primarily on progress with 1) Latent Semantic Analysis as an automatic means of assessing discourse content, 2) PRONET as a sequential method, and 3) CHUMS, a sequential method under development.

The data for these analyses were collected from an experiment, in which 11 teams of three members flew a simulated uninhabited air vehicle for 10 missions (Cooke, Kiekel, & Helm, 2001). Each member had a specialized role. All actions of the team members were recorded and their speech during the first seven missions was transcribed.

Latent Semantic Analysis

Latent Semantic Analysis (LSA; Landauer, Foltz & Laham, 1998; Foltz, Kintsch & Landauer, 1998) can be used to code content data. LSA is a computational linguistic technique that can measure the semantic similarity among units of text. Its "knowledge" of the language is based on a semantic model of domain

knowledge acquired through training on a corpus of domain-relevant text. Through a statistical analysis of how words occur across contexts (e.g., paragraphs), LSA generates a high-dimensional semantic space, in which each original word, as well as larger units of text (utterances, paragraphs, documents), are represented as vectors in the space. The derived vectors for words and utterances can be correlated by taking the cosine between their vectors. This permits the matching of texts based on semantic relatedness, rather than direct keyword overlap alone.

PRONET

PRONET (Cooke, Neville, & Rowe, 1996) is a sequential analysis that relies on the network modeling tool, Pathfinder (Schvaneveldt, 1990). Transition probability matrices among a set of nodes (i.e., events) are input to the Pathfinder algorithm, and a network representation of prominent pairwise connections among events is generated. In this way, one can use PRONET to determine what events "typically" follow one another, for a given lag. For instance, if two events typically follow one another, with one arbitrary event interloping them, then PRONET would identify a lag 2 connection. Two events that typically occur simultaneously would yield a lag 0 connection.

PRONET furthermore allows researchers to determine "typical" chains of events with more than two nodes. Longer chains are evaluated by testing each node pair within the chain, at all required lags. For instance, the chain A→B→C→D could only be considered "typical" if transitions are retained at lag 1 for A→B, B→C, and C→D; at lag 2 for A→C and B→D; and at lag 3 for A→D.

Using physical sequential data, we examined "typical" global patterns of turn-taking in speech acts (Kiekel, Cooke, Foltz, & Shope, 2001). Six communication events were defined, as beginning or ending a speech sequence by each team member. For instance, if the pilot begins speaking, regardless of what she says, this is considered the beginning of a speech act, by the pilot. The navigator beginning a speech act is a different event, as is the pilot's cessation of speech.

A hypothetical turn-taking sequence might consist of the pilot beginning an utterance, then finishing it, then the navigator beginning, and the pilot beginning another utterance before the navigator finishes. These five events would define the pilot speaking uninterrupted, but then interrupting the navigator. An infinite number of possible turn-taking sequences is possible, and our goal was to identify prominent patterns.

We recorded the events, and retained the order in which they occurred. The six possible events were represented as nodes in the subsequent analysis. The set of observed events for each team-at-mission were used to define transition probability matrices of various lags. For example, the node sequence "Pilot begins speaking" followed in the next event (i.e. at lag one) by "Pilot ends speaking" is represented as a cell in the lag 1 transition probability matrix. The event occurs with a probability defined as the number of occurrences for that transition, divided by the sum total of all transitions between "Pilot begins speaking" and all six events.

The transition probability matrices served as input to PRONET, so that we could identify prominent event sequences of speech acts. We have so far analyzed the PRONET output in two ways. First, we qualitatively interpreted the network of "typical" transitions for chains of two events separated by one lag. Here we looked at patterns of networks, and looked for changes over missions.

For the second type of analysis, we derived all "typical" chains of arbitrary length, where a chain's length is equal to the number of nodes it includes. Then, for each team-at-mission, we computed summary statistics on the set of chain lengths (i.e. mean chain length, median, maximum, minimum, and sum of all lengths). These values were used to predict an external, composite measure of team performance.

CHUMS

Clustering Hypothesized Underlying Models in Sequence (CHUMS) is a clustering approach to determining discrete pattern shifts in sequential data. It works by clustering putative models defined by segments of the sequential data. Sequential data are first segmented into discrete units. Each unit is used to estimate a model against which to test the data in every other unit. The matrix of model fit values comparing each unit to each other unit is used as a similarity matrix, with adequate fit values indicating similar units.

Assuming there are some adequate fit values, the data from the two units with the best fit are pooled. The set of hypothesized underlying models is then re-estimated with the remaining clusters. This pairwise model testing process continues iteratively until all clusters are statistically distinct from one another (i.e. all pairwise model tests yield inadequate fits). Remaining clusters can be further analyzed in any way of interest to the researcher. The procedure is fully automated by custom software.

One promising application of CHUMS is in counting the number of distinct communication models remaining, after clustering those that have no statistically detectable difference. For this approach, CommLog data were used to define models of communication dominance among team members. We separated the mission into segments of one-minute duration, and formed multinomial models of how much each team member spoke during each minute. Parameters for each model were estimated by the proportion of speech exhibited by each person, relative to total possible speech for that minute.

These models were then used to generate expected

values for every other minute, whose observed deviation was tested with a Pearson's χ^2. CHUMS yielded counts of distinct models the team exhibited during each mission.

The number of statistically distinct communication models in any mission can be thought of as a measure of communication stability for that mission. This was used to measure how well the team had an established process for the passing of information. The more patterns exhibited during a mission, the less stable the team's communication, and so the less stable their predicted team cognition.

If this is a good measure of the stability of team cognition, then it should be related to performance and situation awareness. For each team-at-mission, we predicted external objective measures of performance and situation awareness from (a) the number of distinct models and (b) models per minute (i.e. the number of models divided by the number of minutes in the mission).

RESULTS AND DISCUSSION

Results Relevant to Physical Data: PRONET

One qualitative outcome of our use of PRONET as described above has been to identify sudden changes in the communication pattern between missions. Curious changes in turn-taking can suggest hypotheses about team process. Analyses pertaining to this approach are discussed more fully by Kiekel, Cooke, Foltz, and Shope (2001).

Figure 1 shows a representation of a PRONET-derived network for lag 1 event transitions. Arrows indicate sequence, such that if the node "Abeg" points to "Aend," it means that person A beginning a speech act is "typically" followed by person A ending a speech act.

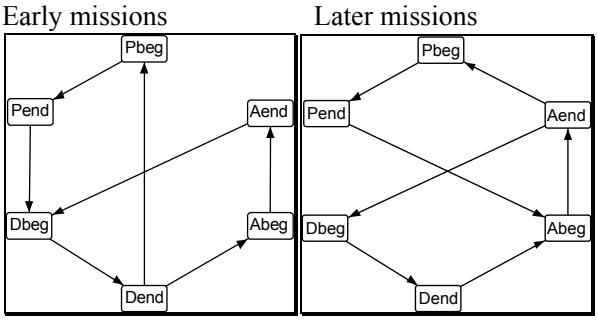

Figure 1. PRONET representation of shift in lag-1 team communication pattern across missions.

For this particular team, the first four missions showed a pattern such as that on the left, and for the last six missions showed a pattern similar to that on the right. The left network we will call "D centered," because person D is the hub of the network. The right network we will call "A centered," because person A is its hub. Here, to be the hub of the network means that (a) the center person tends to finish after they begin speaking, and (b) they tend to begin a speech act after both of their partners finish their speech acts.

The meaning of these networks must be interpreted with caution, as they only represent chains of one event, i.e. the transition at lag 1. The benefit of only looking at chains of length 1 is that they can easily be graphically displayed. The drawback is that the multiple step paths that the image seems to imply are not, in fact implied.

We have gone on to identify the set of all paths of arbitrary length for each team-at-mission. For this, we employed the more quantitative method of recording summary statistics on the set of all "typical" chains for each team-at-mission. These variables have shown some promise for predicting performance. Three promising variables are the minimum chain length, the maximum length, and the median length at each mission.

A multiple regression using some combination of Minimum, Median, and Maximum yields adequate predictions for missions 2, 3, and 5. For mission 2, a model with Maximum and Median yielded $R^2 = .509$, $F(2, 8) = 4.144$, $p = .058$ ($\beta_{max} = -.800$, $\beta_{med} = .913$). For mission 3, a simple linear regression with Minimum yielded $R^2 = .275$, $F(1, 9) = 3.415$, $p = .098$ ($\beta_{min} = -.524$). For mission 5, model with Minimum and Maximum yielded $R^2 = .628$, $F(2, 8) = 5.074$, $p = .051$ ($\beta_{min} = 1.437$; $\beta_{max} = -1.117$).

Collectively, these variables can be thought of as a measure of the team's consistency in turn-taking behavior. They are useful predictors of performance, primarily during the early missions, when skill acquisition is still underway.

Other uses for PRONET are under current investigation. For instance, using PRONET, researchers may be able to predict team performance form the mere presence of repeated, dramatic changes in turn-taking. For instance, if each session has one person at the center of the network, but there is no pattern as to who that center person is, then it may be that the team's communication pattern is unstable. This should be reflected in a low performance score. This method has not yet been tested.

A somewhat more quantitative variant on this approach is to look for particular patterns that predict good performance in teams. For instance, in this task, it appears that the best teams show non-navigator-centered networks for early missions (when they are learning the task), but switch to navigator-centered networks later on. These results are not reported in detail, because they are still being examined as of this writing.

Results Relevant to Physical Data: CHUMS

We now turn to the use of CHUMS to predict external measures of performance and situation awareness from the number of distinct models, and models-per-minute. When analyzed on a team-by-team basis, the success of this method has been mixed so far. There have been very good predictions for some teams, and very poor predictions for others. The search for mediating variables

is underway.

But it is of greater interest to analyze these data mission-by-mission, across teams as our random variable. Mission-by-mission analyses yield negative correlations with performance and situation awareness (SA), as hypothesized. If the number of patterns a team tries out is a measure of cognitive instability, then more models (and models per minute) should lead to poorer performance and SA.

In predicting performance, tests on models per minute showed nothing very impressive. The best prediction was for mission 2, with $R^2 = .243$, $F(1,8) = 2.564$, $p = .148$. While this accounts for over 24% of the variance in performance, it does have a big p-value.

Predicting performance with number of models was better. For mission 2, we saw $R^2 = .373$, $F(1, 8) = 4.753$, $p = .061$. Mission 7 yielded $R^2 = .287$, $F(1, 9) = 3.631$, $p = .089$. In the honorable mention category, we have mission 3 ($R^2 = .212$, $F(1, 8) = 2.151$, $p = .181$) and mission 8 ($R^2 = .255$, $F(1, 8) = 2.731$, $p = .137$.

In attempting to predict SA, the only good prediction was for mission 7, where models per minute ($R^2 = .426$, $F(1, 8) = 5.934$, $p = .041$) was a better predictor than number of models ($R^2 = .384$, $F(1, 8) = 4.982$, $p = .056$). Prediction was fairly good for mission 9 for models per minute ($R^2 = .290$, $F(1, 8) = 3.275$, $p = .108$), but the correlation was in the wrong direction. We are at a loss to explain this baffling mystery.

It does appear, generally, that measures of team communication consistency are more predictive of performance during the learning acquisition phase of a task, and more predictive of SA after asymptote.

Other potential (but as yet untested) applications of CHUMS lie in specifically identifying which patterns remain after clustering. For instance, does the pilot usually speak the most during the first half of the mission, then suddenly clam up when workload gets high? Specific patterns could potentially be examined for flow between minutes, using sequential methods.

Another use of CHUMS that we have not yet begun to investigate is in using models other than the multinomial-team-member-dominance model. For instance, there is no reason not to define a regression equation as the model under consideration. One could model the linear relationship between two speakers at each minute, and see if the pattern changes over time. Any model can be used, as long as it can be tested and assigned a model fit value for a data set.

Results Relevant To Content Data: LSA

LSA is a versatile tool, having been applied to a range of types of written and spoken discourse In the present study, the simplest uses of LSA have been to analyze vector length, sentence-to-sentence coherence, and vector variance. These techniques each have mixed levels of usefulness in the present context. We used the length of the vector representing each utterance as a measure of the amount of task-related information being conveyed.

The ratio of average LSA vector length per mission to average number of words per utterance for a mission (i.e., static communications efficiency), predicts a negative quadratic relationship between communications efficiency and mission performance ($t(48) = -2.5$, $p = .016$). This usage of LSA defines an optimal level of information transmission per communication that is descriptive of top performances.

A somewhat less automatic, but more content-oriented use for LSA is to extract communication content codes. This can be done by using LSA to generate a correlation matrix of utterances with other utterances, then cluster highly correlated utterances. Each original utterance can be classified according to the cluster of which it is a part. Results of human-human reliability for tagging have a c-value of 0.73, while LSA-human reliability is at 0.57. This method, combined with PRONET, has been used to identify typical communication content sequences among teams (Kiekel, Cooke, Foltz, & Shope, 2001). Frequency of occurrence for a set of common content sequences can then presumably be used to predict performance. This work is currently underway.

LSA has also been used to predict overall team performance scores by correlating entire mission transcripts with one another. For this approach to be useful, it requires that a subset of transcripts be associated with known performance scores. The score of a "new" transcript (i.e. one with unknown performance score) can be estimated by computing its proximity to all the "known" transcripts. A proximity-weighted average of the 10 closest "known" transcripts' performance scores along with other statistical measures of the content are taken to be the predicted score for the "new" transcript.

After accounting for the repeated measures structure of the same 11 teams over 7 missions, we found a positive correlation between the LSA measure and performance, $R^2 = .389$, $F(1, 49) = 31.217$, $p < .001$.

Mission by mission analysis shows that this measure is more predictive for earlier missions than for later. Mission 1 yields $R^2 = .707$, $F(1, 8) = 19.269$, $p = .002$. Mission 2 yields $R^2 = .558$, $F(1, 8) = 10.093$, $p = .013$. Mission 3 yields $R^2 = .709$, $F(1, 8) = 19.515$, $p = .002$. Mission 4 yields $R^2 = .290$, $F(1, 8) = 3.275$, $p = .108$. Mission 5 yields $R^2 = .626$, $F(1, 7) = 11.714$, $p = .011$. The latter two missions did not yield adequate predictions. This may be due to there being greater inter-team differences in the earlier missions.

Other approaches to using LSA for team cognition are also under evaluation. One way to use LSA to measure coherence is to take the mean cosine between each utterance and its sequel. This measure shows a U-shaped curve in its relation to performance, with either too little or too much coherence leading to poorer performance (Foltz,

Kintsch & Landauer, 1998).

Another untested candidate for a team cognition measure is the measurement of what information is being conveyed by which team members. This task has a particular ideal information flow pattern. Hence, we would be interested in assessing how far a particular team-at-mission deviates from that ideal. The approach would be to specify *a priori* which actual content should be appropriately conveyed by which speakers. This can be done by creating excerpts that are prototypically representative of the information in question. Then researchers can classify actual team utterances by correlating each utterance with each excerpt. This tells us approximately what each person is saying when they speak.

CONCLUSIONS

The methods presented here can be manipulated in an infinite number of ways. We have only begun to examine the potential uses. Once we have derived a set of known valuable methods, it will be possible to create software and hardware that can very rapidly collect and analyze these data.

Since many of the methods require no human intervention, the analyses can be conducted as fast as the machinery that processes it can handle. Such tools have the potential to be used to characterize team performance in real time. Decisions made in real time can be used for interventions of various sorts.

These preliminary results show strong promise in using automated communication methods to measure team performance and cognition. Most of the methods are predictive of performance. Tools such as these permit a holistic measurement of team cognition, while avoiding some of the pitfalls of time consumption and weak reliability.

ACKNOWLEDGEMENTS

This work was supported by ONR Grant No. N00014-00-1-0818 and greatly benefited from contributions of Steven Shope and Susan Stevens.

REFERENCES

Cooke, N. J., Rivera, K., Shope, S.M., & Caukwell, S. (1999). A synthetic task environment for team cognition research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting,* 303-307.

Cooke, N. J., Kiekel, P. A., & Helm E. (2001). Measuring Team Knowledge During Skill Acquisition of a Complex Task. *International Journal of Cognitive Ergonomics: Special Section on Knowledge Acquisition, 5,* 297-315.

Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996). Procedural network representations of sequential data. *Human-Computer Interaction, 11*, 29-68.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42*, 151-173.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25(2 & 3),* 285-307.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., & Shope, S. M. (2001). Automating measurement of team cognition through analysis of communication data. In M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek (Eds.), *Usability Evaluation and Interface Design*, pp. 1382-1386, Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2&3), 259-284.

Schvaneveldt, R. W. (Ed.) (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.