Non-Intrusive Detection of Psycho-Social Dimensions using Sociolinguistics

Peggy Wu, Jeffrey Rye, Christopher Miller, Sonja Schmer-Galunder, Tammy Ott
SIFT (Smart Information Flow Technologies)
Minneapolis, MN
{PWu, JRye, CMiller, SGalunder, TOtt)@siftech.com

***Abstract*—Long duration space flights such as a two and a half year mission to Mars present many unique challenges to the behavioral health of astronauts. Factors such as social monotony, workload, a confined environment, sensory deprivation, and limited access to family and psychosocial support can affect crew welfare and task performance. NASA recognizes a "risk of performance decrements due to inadequate cooperation, coordination, communication, and psychosocial adaptation within a team;" reports from Mir revealed that conflicts between crew members have resulted in early termination of missions. Currently, flight crews and support staff have real time voice and video communications capabilities on the International Space Station to keep astronauts connected, and allow operations staff to monitor the crew's well-being. However, communications for long duration missions will likely be limited and disrupted by time latencies. Crew workload may also prohibit crew members from providing the extensive self-reports that the Earth-bound support team needs to accurately access the crew's psychological health. Further, the metrics of interest are difficult to obtain because some are inherently qualitative, while others may not be amendable to self-reports. We first describe an extensive review of psycho-social dimensions relevant to long duration space flight, their manifestations, and possible detection methods. We then describe a novel method of non-intrusive detection developed initially for application in the Empire Challenge military exercise in 2010. This system, called ADMIRE for Assessment of Discourse Media Indicators of Relative Esteem, leverages prior work in cultural and socio-linguistic theory to develop standardized, non-intrusive methods for data collection and knowledge extraction about factors salient to group psychosocial dynamics. Finally, we describe our approach to follow-up work applying ADMIRE to historical space flight data, as well as in ongoing studies in space analog environments to identify potential changes in individual and team psycho-social factors before they lead to deficits in health and task performance.**

Keywords—Long Duration Space Mission, Psycho-social Measures; Socio-linguistics; Politeness; Power and Social Dynamics; Communications; Non-intrusive Detection

I. INTRODUCTION

Work on the International Space Station (ISS) is performed by culturally-mixed, international teams [1], with team members potentially never having trained together [2]. This is also likely to be true of any deep-space, long duration missions such as manned flight to Mars. Prior research (e.g., [3,4]) has shown that diversity within teams can have positive impacts on performance—particularly because it offers a variety of perspectives and skillsets. However, others (e.g., [5,6]) have found that decrements in communications and decision making assumptions, whether stemming from cultural or training backgrounds, can defeat any benefits from diversity and provoke team breakdowns and sub-optimal performance. Our own work has shown that simply feeling like being a part of a team ("affiliation") can have positive effects on performance and directive compliance, and that affiliation effects can be influenced by verbal behaviors [7]. Our most current work shows that a computational approach to inspecting verbal interaction behaviors can provide insight into team cohesion and affective relationships [8].

As a part of its Human Research Program Integrated Research Plan (IRP), NASA has asked researchers to identify "the most likely and serious threats to task performance, teamwork, and psycho-social performance" during long duration flight. Such threats stem from individual and cultural differences in perceptions and expectations regarding a diverse set of factors ranging from leadership and team dynamics, to workload, to meaningful work. While substantial efforts have been made to identify psychosocial topics relevant to long duration space flight (see [9]), there is conflicting evidence about the effects of some factors on performance (e.g. conscientiousness, see [9], p 18). Such research is plagued by three nearly universal problems:

1. Collecting naturally occurring behavioral data and personal perceptions (e.g. interpersonal relationships, affect, etc.) is difficult without intruding upon, interrupting or otherwise distorting individuals' behaviors.
2. Extracting meaningful data from lengthy records can be a laborious process subject to the opinions and the cultural and individual biases of the data analysts.
3. Validating the results of various measures is itself a fundamental problem.

For these reasons, it is necessary to use repeatable, non-intrusive, valid and efficient assessment methods to detect salient states of crew psychosocial wellbeing and performance. At the same time, several rich sources of data that are amendable to linguistic analysis for deriving psycho-social states in a non or minimally intrusive fashion currently exist. In a prior study, Stuster [10] showed the feasibility and

ASONAM 2013, August 25-28, 2013, Niagara Falls, Canada
978-1-4799-1498-2/13/$31.00 ©2013 IEEE

analytic value of assessing personal journals from astronauts on the ISS. Similarly, the authors [8] have demonstrated the use of correspondences and communications such as instant messages and transcribed telephone transcripts to derive power measures. By creating automated methods to analyze personal journals and naturally-occurring communications behaviors, it is possible to infer a number of behavioral health factors that are otherwise not computationally tractable. The long term goal of our research, then, is to develop a system that can automatically detect changes in the psychosocial dimensions that affect individual and team behavioral and task performance.

Many individual and team psychosocial and performance precursors manifest themselves in the communication between team members, or in individual written reflections. However, the question remains as to which factors can be detected by automated methods so that we may define the scope and capabilities of the resulting system. In order to systematically identify which dimensions are most amendable to automated detection, we first conducted an extensive review of psycho-social dimensions of interest and prior automated and non-automated methods. We provide a description of this survey below.

II. A SURVEY OF RELEVANT PSYCHO-SOCIAL FACTORS

A. Objectives

Our ultimate objective is to create technology that can unobtrusively monitor individual and team psycho-social health dimensions that may have the most impact, either positively or negatively, on team members' health and performance. There are many existing studies on the psychology and behavioral health issues in ICEs (Isolated Confined Environments), but few focus on automated detection methods. Overlapping concepts and terminology in the literature led us to organize and conduct our review in the following fashion. We:

1) reviewed prior work that used an evidence-based approach to identify psycho-social dimensions that affect individual or team health and performance in spaceflight or space analogs

2) identified possible observable manifestations of psycho-social dimensions, and

3) selectively matched those observable manifestations with current or near term technologies and analytical methods that are non- or minimally-intrusive to flight and ground crew members.

We placed several inclusion limitations on psycho-social factors. While there are behavioral risks throughout the entire mission, for our purposes, the most critical timeframe is in-mission as opposed to the pre- or post-flight due to the anticipated reduction of in-flight communications capacity. We also focused on team-related dimensions, as a future mission will likely use a small crew as opposed to a single individual. However, we did include some factors observed at the individual level, as an individual's psychological state will undoubtedly affect a team's state. For many of the psycho-social dimensions, we list manifestations that are likely to be detected by current or near term technologies. We concurrently surveyed detection methods, especially those that are non-intrusive or minimally intrusive, and matched them to detect manifestations of psycho-social dimensions. The concept and basic organization of the survey is shown in Figure 1.

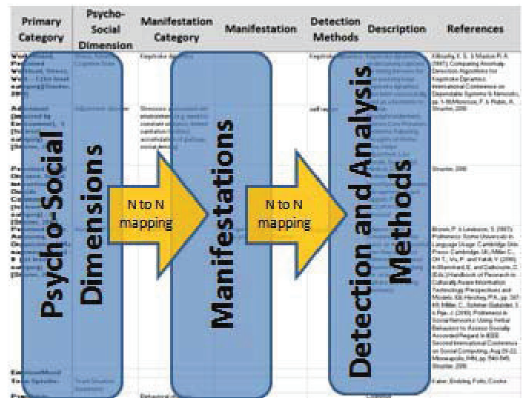

Figure 1 Concept of Psycho-social Dimensions Survey

B. The Psycho-Social Dimensions and Detection Survey

We reviewed over 50 publications and technical reports to extract possible psycho-social dimensions of interest. The resulting list contains over 85 psycho-social functions that are neither orthogonal nor independent, though we grouped similar items and concepts together wherever possible, and organized them into top tier categories. To provide a sense of priority, we ranked the categories roughly based on their likely importance to space operations as guided by the frequency with which issues related to those categories were reported in the Stuster astronaut journal study [10], as well as the general indication of their importance from the NASA Behavioral Health and Performance evidence book [9]. The survey is available on a case-by-case basis through contact with the author.

The top tier categories, in ranked order, are:

- Work, Perceived Workload, Work related stress
- Adjustment (imposed by Environment)
- Perceived Social Distance, Social Interaction
- Perceived Power, Organizational/Management related
- Emotion/Mood/Mood Disorders
- Team Specific
- Psychosomatic Reactions
- Gender and Cultural Differences
- Salutogenic (i.e. positive, motivating, awe-inspiring) Experiences
- Individual Traits and Personality
- Others

The next step was to associate observable manifestations with the identified psycho-social dimensions. Many of them had multiple and overlapping manifestations. In general, the manifestations fell into the following categories:

- Individual Behavioral Changes (e.g. changes in social interaction, helping behaviors, voluntary cooperation, temperament, hallucinations, motivation, daily tasks reported as monotonous, creativity, affect, keyboard dynamics, altruism)
- Team Related Behavioral Changes (e.g. changes in coordination style such as more time-consuming, explicit coordination strategies vs. streamlined implicitly coordination to shared mental models and information, displacement, crew autonomy)
- Non verbal communication (e.g. body posture, facial expression, intonations, proximity)
- Linguistics (e.g. use of filler words, politeness, word frequency, positive/negative emotion words)
- Physiological (e.g. gaze direction, heart rate)
- General Health, Food Intake, Sleep (e.g. quality and quantity, sleep hygiene, sleep disturbances)

The last step in this process was to identify detection and analytical tools that can be used for sensing the psycho-social manifestations. During this process, we considered the environment and context in which data is gathered so that the results would be meaningful and useful. For example, while proximity between individuals can indicate social distance [11], flight crew members on the ISS often communicate using the onboard communications system, thus rendering the physical proximity meaningless in predicting social familiarity. Microgravity, as well as the confined habitat in a notional spacecraft, may also distort the usefulness of analyzing the body language and physical orientation of individuals. In our review, we found that by far, current methods of detecting manifestations rely on interviews and surveys. Other reporting methods are emerging, such as assessment methods in the form of games, with non-intrusive and minimally intrusive methods making up a small minority. We organized the detection methods into the following types:

- Linguistic Analysis
- Non verbal communication Analysis (e.g. Optical Computer Recognition for facial expression, Keystroke dynamics, proxemics)
- Physiological and Neuro-physiological monitoring (e.g. voice intonation, socio-meter)
- Performance Measures (e.g. reaction time)
- Environmental sensors
- Sleep monitoring (e.g. Actiwatch)
- Self Report in the form of games
- Self Report/Survey/Observations

The result is a live spreadsheet containing over 85 psycho-social factors with an n-to-n mapping to over 80 manifestations of various levels of granularity. Some of the manifestations, in turn, are mapped to some of the eight aforementioned detection categories. Through a series of prior projects, we have focused on a small subset of psycho-social dimensions and their manifestations. Below, we discuss the availability of data and our detection methods to derive relevant psycho-social dimensions.

C. *Available Data and Psycho-social Dimensions of Interest*

Under our NASA-funded effort to explore behavioral health and performance in long-duration missions, our objective is to systematically identify and test a select number of non-intrusive methods of detecting factors that can affect individual and team behavioral health. To this end, we are using the psycho-social dimensions survey to identify manifestations that can be detected in non-intrusive and minimally-intrusive methods. For reasons ranging from astronaut privacy, to the high workload that limits the crew's ability to complete surveys in-flight, to the cost of deploying new sensors on the ISS, available data and data collection methods are highly limited. Thus, we are examining non-intrusive detection methods based on "data" collected by existing "sensors" (including, especially, existing communications streams) and further distinguishing related manifestations that can inform us about the maximum number of relevant psycho-social dimensions as identified in our survey. This process produced a limited set of data types, which mainly include flight crew communications, flight to ground communications, and astronaut journals (including blogs and tweets). Fortunately, due to prior work, we have a small arsenal of analyses and detection methods to use against these data types. Further, the individual and group dimensions that can be derived from available data streams have been shown to affect task performance and are relevant at both the individual and group level. Below, we describe these dimensions in greater detail and provide their theoretical background.

D. *A Theory of Group Dynamics Expressed Through Politeness*

A large proportion of team performance and psychosocial problems—and problem precursors—manifest themselves in the communications between team members—whether within flight crews or between crews and the ground. Communication (both verbal and non-verbal) is, after all, how humans detect and manage interpersonal and team cohesion issues. Communication, in all its nuances, may well take on even more importance for long duration space flights [10], especially under the circumstances of social monotony, possible discrepancies in cultural interpretations, and delayed communication with ground crews. In previous work, we developed a quantitative, computational implementation of the Brown and Levinson [12] model of "politeness" in discourse and have demonstrated [13] its ability to predict perceptions of politeness in members of various cultures, and to generate polite discourse for use in virtual agents that were perceived as believable by native speakers.

Brown and Levinson [12] theorize that politeness behaviors allow humans to signal our beliefs about social relationships and attempts to manipulate the beliefs of others. Further, their cross-cultural empirical studies suggest that the use of politeness constructs are universal, even though their manifestations are culture specific. In our interpretation of Brown and Levinson's model (cf. Fig. 2 and [14]), all social interaction threatens "face" [15] and the amount of politeness "value" used must offset the threat if the social status quo is to be maintained. In social interactions, the degree of face threat is a function of the power and social distance (roughly, familiarity) of the participants and the degree of imposition of the interaction. To offset face threat, we make use of polite "redressive strategies," which are the familiar verbal and non-verbal behaviors which we think of as "polite": "please," "thank you," honorifics, apologies, etc. These strategies are culture-specific, but Brown and Levinson's data suggest that the broad functional categories of redressive strategies are repeated across cultures. Each type of redressive behavior has a *redressive value* within a culture—meaning that it is capable of offsetting a certain amount of face threat. If a hearer perceives less redressive value being used than the s/he perceives as necessary given the face threat inherent in the interaction, then

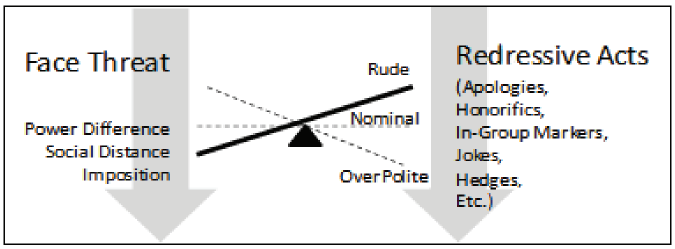

Figure 2. Conceptual depiction of Brown and Levinson's functional model of politeness.

the interaction will be perceived as rude; if more redress is perceived than is deemed necessary, then the interaction will be perceived as "over polite." The fact that the manifestations of politeness strategies are culturally specific means that in implementation, the culture of the actors signal the use of different scores for the redressive strategies (e.g. bowing may have a

The value of both a threat and its redress is based on the perceptions of an individual, which explains how the same utterance can be polite or rude depending on context. If I use the same phrase to ask you to pass the salt (e.g., "Excuse me sir, but would you mind passing the salt?"), I might be perceived as being too polite if we are close family members (low social distance), or I might be a bit rude if you are the President of the United States (high power difference)—or if you are engaged in an important activity such as delivering a speech (high imposition). Similarly, this model can explain how one individual can intend an act as nominal while another can expereince it as rude—if there is a mismatch between their perceptions of any of the factors used to determine the degree of face threat, or in their perceptions of the redressive value of the politeness behaviors used

A final aspect of our interpretation of the Brown and Levinson model is that politeness perception is a cognitive process. Our perceptions of social relationships and the imposition of an act may start off as nominal, but we draw inferences from the transaction itself. When we perceive an imbalance, we try to explain it by adjusting our assumptions about the relationship or imposition—or the character and knowledge of the participants. This is why if you burst into my office and yell "Get out now!" I am more likely to believe that your language was due to an urgent (imposing) danger than that you're just rude. However, if you persist in this behavior, or I learn that there is no danger, I may adjust my perceptions of your actions and attempt to find alternate interpretations (e.g., perhaps you are trying to assert higher power over me). In the next section, we describe our use of this model to automatically derive a depiction of the power dynamics within a group using computerized recognizers to identify and score politeness behaviors in a military chat context.

III. THE ADMIRE APPROACH FOR ANALYSIS OF COMMUNICATIONS DATA

In prior work funded by DARPA [14], we demonstrated the use of the Brown and Levinson theory as the basis for a computational model of perceived politeness. We have shown that, in American English cultural examples, our implementation is reliable at predicting perceived politeness for both trained and naïve raters. We have also demonstrated the ability to "swap" knowledge bases of culture-specific rating values and provide both predictions and explanations for alternate cultural perceptions of interactions (in American English, Pasthun and Modern Standard Arabic). Finally, we have used our computational model in language training games and rehearsal aids, demonstrating exponential scale up savings in software engineering costs.

In more recent work funded by the U.S. Navy's Office of Naval Research, we used the politeness model with observed behaviors between individuals in a social network to infer their social relationships and attitudes toward each other. In principle, this is what an outside observer does naturally. If I see person A calling person B by an honorific, using formal language, making requests instead of demands, etc., I am likely to infer that person A has (or, more accurately, believes themself to have) less power than Person B, that A and B are not close friends and/or that A is imposing on B to a degree. The challenge of our program was to enable a software algorithm to make similar inferences in an automated fashion.

Under a program called ADMIRE (for Assessment of Discourse Media Indicators of Relative Esteem), we developed a parsing, recognition and scoring system that can analyze a corpus of textual data in any of various formats (e.g. email, IM Chat, discussion forum/blog, transcribed phone messages). The corpus was processed to identify speakers and addressees and parse out potentially misleading aspects (such as quoted text from prior emails). The resulting set of interaction exchanges was run through a recognition and scoring algorithm based on our computational implementation of the Brown and Levinson model to recognize and score specific instances of politeness behavior usages indicative of particular relationships between interactants. In the subsections below, we first describe the ADMIRE implementation, and then the results of applying ADMIRE to a series of military chatroom exercises.

A. The ADMIRE Architecture and Implementation

Our ADMIRE tool was developed to infer relationships (we have primarily investigated power relationships, but others are possible) expressed in interactions between individuals (ranging from dyads to, potentially, hundreds of individuals) and captured in textual records of those interactions.

1) Functional Flow

ADMIRE processes a corpus of textual discourses in order to generate a set of output files describing the relationships between actors in the corpus. ADMIRE accomplishes this in the following way; it:

1. parses the corpus files to create a set of discourses
2. populates a communications network from the discourses and filters it for efficiency
3. identifies uses of politeness strategies in the remaining discourses
4. computes a score summarizing how much redressive "politeness" (as a positive or negative score centered on zero for a neutral act) was directed from each speaker to all hearers in each discourse
5. ranks the actors to derive an overall ordering,
6. writes output files describing the results

Below, we will discuss each of these steps in detail.

2) Parsing and Data Representation

ADMIRE represents interactions as a directed graph with actors as nodes and communications as unidirectional edges. Where two actors speak to each other, ADMIRE represents the conversation as a pair of links. When a discourse includes three or more speakers, the network contains links from each speaker to each hearer.

ADMIRE derives the communication network from a series of discourse interactions, where a discourse represents a block of communications from the corpus. ADMIRE uses specialized discourse representations based on the nature of the corpus. For emails, ADMIRE lists the actors in the To: and CC: fields separately and stores the Subject: field. During parsing, ADMIRE eliminates contents being forwarded or replied to, so the contents of an email represent only what was written by the sender. When asked for the set of hearers, emails return the union of the To: and CC: fields. When asked for the speaker, emails return the Sender:.

In contrast, ADMIRE uses a dialog specialization of discourse to represent spoken dialog (e.g., from transcripts) and chat (IRC) transcripts. Dialog instances are conversations with multiple speakers and hearers. Unlike emails, dialog instances may return multiple speakers. A dialog generally represents a discrete conversation between small numbers of people. When dialogs have too many actors, ADMIRE attributes redressive acts too widely, thus reducing analysis effectiveness. The hearers for a dialog are generally the same as the speakers, but may include actors who never speak.

ADMIRE transforms corpora consisting of flat text files into Discourses using an extensible parser. We created different parsing adapters for email, IRC chat, phone logs and, under NASA funding, radio communications logs. Each comes with different method for processing sender and recipient identification and for identifying novel content vs. quoted prior content.

Email clearly defines the speaker and hearer(s) in the header. Phone transcripts typically contain two or three actors, allowing ADMIRE to assume that each line of content is addressed to everyone in the dialog. IRC, however, poses a challenge in revealing who is being addressed with each statement. To address this, ADMIRE has implemented the following three heuristics.

1. *Assume continuity*—If the speaker for a line was the speaker in the previous line, set the current hearer(s) to be the same as the hearer(s) from the previous line.
2. *Assume reciprocity*—If the speaker for a line was a hearer for the previous line, set the current hearer to be the speaker of the previous line.
3. *Assume sequential turn taking*—If the hearer for the previous line was unknown, set the hearer to be the speaker of the previous line.

To assess email and spoken dialog, ADMIRE creates a single discourse for each input file. However, IRC logs can span hours or days and can contain numerous conversations. The adapter segments the transcript into discrete dialogs. As with inferring actors, this is a heuristic process. For each statement, the adapter starts a new dialog if neither the speaker nor the hearer is already in the lists of speakers and hearers for the current dialog. In other words, the existing dialog is reused as long as new statements are spoken or heard by someone already in the dialog.

3) Network Creation and Filtering

ADMIRE builds a network representing who talks to whom and with what frequency differently for emails than for transcripts or IRC. For email data, it adds links from the sender to each recipient in the To: and CC: fields. For dialogs (spoken or IRC), it adds links from each sender to each hearer, as all speakers are assumed to be speaking to everyone else.

ADMIRE can remove unwanted links to reduce the size or complexity of a network. For example, it can remove links involving participants who communicate very rarely (and therefore, for whom we did not have enough data to create reasonable estimated power scores). The filter works by repeatedly applying a set of filtering rules to a network until they stabilize.

ADMIRE defines two filters to reduce network size and to ensure that only significant links are included in the network. These filters can 1) limit the minimum number of discourses that must be on a link, or 2) require links to be symmetrical. We also define two filters to increase confidence of social relationships by requiring a minimum number of discourses. An additional filter enforces that each link uses one of the actors as speaker or hearer. This last filter ensures that the network contains a specific, inclusive subgroup of actors, or that the network must contain the actors and anyone they are linked to.

4) Redressive Act Identification

ADMIRE includes a politeness assessor which analyses discourse contents for the presence of politeness strategies. Strategies are not necessarily polite or rude in the naive sense. Instead, they are speech patterns that reflect a speaker's perception about a social relationship. The politeness assessor identifies sstrategy usage a single discourse at a time for each speaker in each discourse across an corpus.

ADMIRE's politeness assessor uses a set of politeness behavior recognizers primarily based on regular expressions and natural language processing. Examples of recognizers include:

- **Please**. "Please," "plz," or "pls."
- **Gratitude**. "thanks," "thank you," "thankx," etc.
- **LeadingDirective**. LeadingDirectives occur when an utterance begins with an imperative. Within tactical military chat, sample imperatives include "follow," "stay," "move," "report," "standby," "start," and "leave."
- **OutGroup and InGroup**. The OutGroup and InGroup strategies are complementary. OutGroup conveys that the speaker considers the hearer to be in a different group and InGroup suggests that they are in a common group. ADMIRE identifies OutGroup by the presence of "you," "you're," "your," "yours," or similar forms. InGroup is signaled by use of first-person plural terms, such as "we," "us," "our," and "ours." These identifiers are based on the trend cited in [16, 17]. This approach yields false-positives when the speaker is speaking on behalf of the group. Use of the OutGroup strategy suggests that the speaker has power over the hearer, and thus contributes negatively to the power of the hearer, while InGroup does the reverse.
- **Wilco**. In military jargon, "wilco" is an abbreviation of "I understand and will comply." The Wilco strategy is signaled by "wilco" or "will do."
- **ThirdPerson**. ADMIRE identifies ThirdPerson when a speaker refers to him or herself in the third-person.

5) Scoring

ADMIRE uses a configurable scoring model to compute a single numeric value for each speaker in each discourse that reflects the speaker's perception of their relationship to the hearers as derived from their use of politeness strategies. The scoring model is capable of computing various dimensions of social relationships, including: power, social distance, and affect (likability), although to date we have focused only on power relationships.

To compute a score for a speaker in a discourse, ADMIRE examines a number of factors, including the number of sentences and number of words they speak. However, the main component of the score is based on the presence and sometimes absence of specific politeness strategies, weighted by the number of opportunities for the strategy to be used.

Within an interaction medium and culture, different strategies consistently align with relationship dimensions. However, across domains, strategies may align with different dimensions and may be more or less relevant. For instance, the use of "please" or language implying gratitude in military tactical chat distinctly suggests that the speaker has power over the hearers. However, in spoken dialog, the use of "please" or gratitude language weakly suggests the opposite (that the speaker has less power than the hearers).

6) Ranking

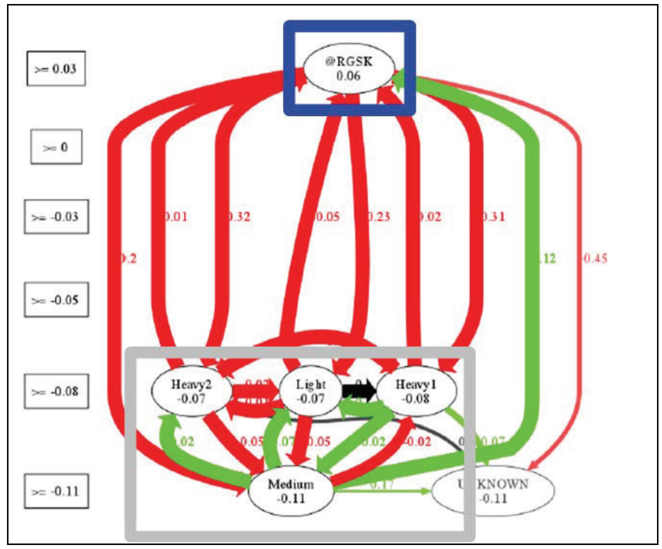

Figure 3. Network graph of power scores and communications relationships derived for one EC10 chat room over the full week's vignettes.

ADMIRE uses a force-directed algorithm to compute overall rank scores for the actors in the network and uses these scores to position nodes in a graph for visualizing the social network. It is important to note that the rankings are meaningful within a corpus of data, but that they are unit-less and are not transferrable across corpora or domains.

ADMIRE's force-directed ranking algorithm uses a physics simulation which treats each link as a spring. The stiffness of the spring is based on the number of discourses, and its equilibrium distance is based on the mean politeness score. Links with more messages are more stiff, and thus have a greater effect on the resulting ranks.

Often, networks will have a positive link from actor A to B and a corresponding negative link from B to A. This relationship suggests mutual agreement that one is higher than the other. In some cases, both links represent relationships with the same sign. If the magnitudes associated with these links are similar, the spring forces will cancel out and other links in the network will govern the ranks of the two actors.

7) Output: Writing

An ADMIRE analysis produces numerous descriptions of the detected social relationships in a corpus. One of the most intuitive outputs is the network graph (an example from our experiment is depicted in Figure 3). ADMIRE positions actors with higher ranks higher on the graph. The graph connects ovals with arrows that show what speakers (tails of arrows) communicated with what hearers (heads of arrows). The width of the arrow indicates how many messages were observed, though width is limited to ensure readability. Arrows are labeled with the mean power accorded from the speaker to the hearer. Higher values indicate that the speaker accords more power to the hearer. ADMIRE colors arrows with positive "power accordance" green, negative accordance red, and neutral (0.0) black. The ranking algorithm chooses ranks such that most positive arrows will point up the page and negative arrows will point down. This same information can also be

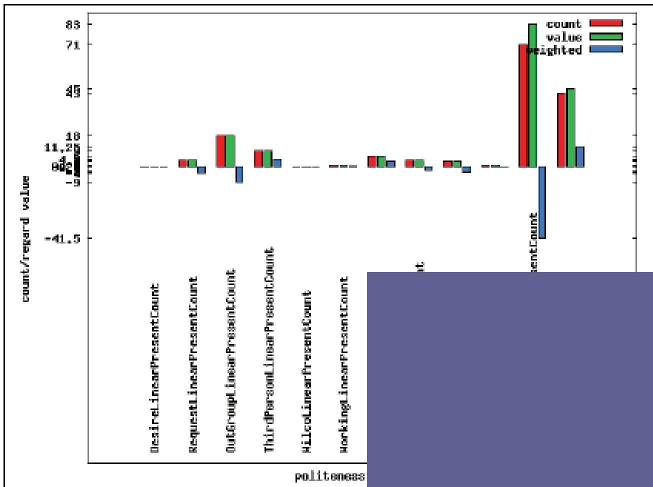

Figure 4. Count, Value and Weight of Po
one of the EC10 ch

delivered as a Comma-Separated Value (CSV) file along with summary values showing the minimum, mean, maximum, and standard deviation for the regard shown in messages.

ADMIRE also provides a graph showing how each strategy contributes to the scoring (see Figure 4). In the graph, weighted values with larger magnitudes indicate that a strategy contributed more to the overall rankings of each hearer and/or speaker. Positive values reflect strategies indicating the hearer (recipient) has more power, and negative values indicate the speaker has more power. This can be particularly useful in identifying the types of politeness strategies most commonly used in a corpus, a specific domain, or culture of interest.

B. Validating ADMIRE Results

During development, we have explored applying ADMIRE to multiple discourse types to deduce power relationships. These included: emails (both our own internal corporate emails and the ENRON email dataset [18], transcribed spoken interactions (the Family Secrets mafia phone intercepts [19]), and IRC chat (two sets of military chat). The most fully realized experiment that we have performed to date is presented below.

1) Experimental Environment: Empire Challenge 2010

In 2010, under sponsorship from the Office of Naval Research (ONR), we developed ADMIRE and participated in the Empire Challenge Green Devil 2 exercise (known as EC10) at Ft. Huachuca, AZ—a joint military exercise in which active duty forces manned advanced sensor and computer equipment to gather intelligence on simulated communities to identify insurgents and forestall attacks. Green Devil 2 was organized and overseen by ONR to test advanced, integrated ISR technologies under realistic field conditions. We provided a demonstration to field ADMIRE, integrating it with ONR's field data architecture to process military tactical chat interactions generated by Marines and support contractors in three different chat rooms during live exercises.

Throughout the exercise, participants coordinated sensor and analysis activities in different IRC chat rooms. Chat messages were generated throughout the week. Each chat room involved anywhere from 8-13 separate vignettes, each lasting 1-2 hours and involving 5-25 individuals. ADMIRE downloaded chat records from 3 chat rooms at the end of each vignette and parsed, scored and created power network graphs as described above.

2) Experimental Results.

Using the recognizers described above, we produced a power score for each participant in each chat room in each vignette. We then qualitatively compared ADMIRE's score against the known or expected "ground-truth" power relationships. Note that while ground truth is often the same as the "org chart" or chain of command, it is quite possible for operations to deviate from it. Figure 4 shows an output network graph for the week's vignettes for one chat room.

	GBOSS	TAC Ops	Unc. GHUB
# Sessions	13	8	8
% Correct Agg.	100%	100%	100%
% Correct Ind. session	82%	94%	83%
Binom. Prob.	p<.00000	p<.00001	p<.00000

Figure 5. Summary results for 3 chat rooms.

Figure 5 summarizes ADMIRE's results over three different chat rooms in the EC10 exercise. When we averaged results over the full week's chat data (13 sessions for one chat room, 8 for each of the two other rooms studied), ADMIRE's end result was 100% successful at detecting the core "ground truth" relationships in each chat room. Even when applied to the individual vignettes (1-2 hours of data), ADMIRE's conclusions about power relationships averaged 84% accurate.

IV. Conclusions and Future Work

The relative success of ADMIRE "in the wild" of EC10, using realistic, military tactical IRC chat generated during highly representative military activities implies that our approach is robust and useful at least with regard to the kinds of task-focused behaviors that were being exhibited there. Additional, prior work, while less formal, has shown promise with spoken dialogue and corporate emails as well.

The polteness recognition strategies we have developed are, we believe, prevalent in general English usage and American cultural domains. Insofar as any other domain has an implicit or explicit power hierarchy, we would expect ADMIRE to have the capacity to detect it, though some tuning of weights and specific jargon might be required.

As described above, in the EC10 exercise, ADMIRE was used solely to detect power relationships. The Brown and Levinson model, described at the beginning of this paper, suggests that politeness behaviors should also be the key to detecting social distance/familiarity as well as the degree of perceived imposition on the hearer in various interactions.

Through our survey of space flight literature, we believe that these dimensions are relevant to understanding individual and team psycho-social health for NASA applications. We have begun working to modify ADMIRE for power relationship detection through non-intrusively analyzing both simulated exercises and transcribed historical data including Apollo mission radio communications to see if we can detect power relationships and shifts in this data. Sudden and unanticipated shifts in power or other social dynamics can indicate potential problems or salutogenic events, and a non-intrusive method of detecting them can keep ground based support staff and flight crews connected while maintaining crew privacy minimizing staff workload.

Acknowledgment

The above work was sponsored by The U.S. Office of Naval Research under contract #N00014-09-C-0264 and NASA's Human Research Program under contract #NNX12AB40G. We would like to thank our ONR Program Managers Dr. Martin Kruger and Ms. Maya Rubeiz for the opportunity to participate in EC10. We would also like to thank our NASA Program Mangers Lauren Leventon, Brandon Vessey, and the BHP Element for their oversight, direction, and support.

References

[1] Boyd, J, Kanas, N, Gushin, V. & Saylor, S. (2007). Cultural differences in patterns of mood states on board the International Space Station *Acta Astronautica, Vol 61* (7-8), 668-671

[2] BHP workshop working notes, 2010

[3] Maznevski, Martha L. (1994). Understanding our differences: Performance in decision-making groups with diverse members. *Human Relations 47*(5). 531-552.

[4] Lindgren, I., & Smith, K. (2006b). National patterns of teamwork during an emergency management simulation. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting.* San Francisco, CA, 354 - 357.

[5] Mateev, A. V. and Nelson, P. E. (2004). Cross Cultural Communication Competence and Multicultural Team Performance: Perceptions of American and Russian Managers. *International Journal of Cross Cultural Management* 4(2): 253–70.

[6] N. Cooke, E. Salas, P. Kiekel, and B. Bell. Advances in measuring team cognition.In E. Salas and S. Fiore, editors, Team Cognition: Understanding the Factors that Drive Process and Performance. American Psychological Association, 2004.

[7] Miller C., Ott T., Wu, P. and Vakili, V. (2010). In Blanchard, E. and Dalhousie, D. (Eds.) *Handbook of Research in Culturally Aware Information Technology: Perspectives and Models.* IGI; Hershey, PA., pp. 387-411.

[8] Miller, C., Schmer-Galunder, S. & Rye, J. (2010). Politeness in Social Networks: Using Verbal Behaviors to Assess Socially-Accorded Regard. In *IEEE Second International Conference on Social Computing*, Aug 20-22, Minneapolis, MN, pp. 540-545.

[9] Mcphee, Jancy C. and Charles, John B. eds. (2009). *Human Health and Performance Risks of Space Exploration Missions.* Accessed Nov 2010: http://humanresearch.jsc.nasa.gov/files/HRP_EvidenceBook_SSP-2009-3405.pdf

[10] Stuster, J (2010). *Behavioral Issues Associated with Long-Duration Space Expeditions: Review and Analysis of Astronaut Journals*. Experiment 01-E104 (Journals): Final Report. Accessed Nov 2010, http://ston.jsc.nasa.gov/collections/TRS/_techrep/TM-2010-216130.pdf

[11] Hall, E.T. (1966). *The Hidden Dimension*. Anchor Books. ISBN 0-385-08476-5.

[12] Brown, P. & Levinson, S. (1987). *Politeness: Some Universals in Language Usage.* Cambridge Univ. Press: Cambridge, UK.

[13] Miller, C. & Wu, P. (2007). The "Etiquette Engine™": A Computational Model of Social Interaction Politeness. In *Proceedings of 51st Conference of HFES.* 1-5 Oct, Baltimore, MD.

[14] C. Miller, P. Wu, H. Funk, L. Johnson, and H. Viljalmsson. A Computational Approach to Etiquette and Politeness: An "Etiquette Engine™" for Cultural Interaction Training. In Proc. 2007 Conference on Behavior Representation In Modeling and Simulation, pp. 26-29, Norfolk, VA; March 26-29, 2007.

[15] E. Goffman. Interactional Ritual. Chicago: Aldine, 1967.

[16] Y. Tausczik, and J. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J of Language and Social Psych, vol. 29, no.1, pp. 24-54, 2010.

[17] J. Pennebaker. The Secret Life of Pronouns. Bloomsbury Press, NY, 2011.

[18] Available at www.cs.cmu.edu/~enron/ on 9/12/12.

[19] Available at www.justice.gov/usao/iln/hot/familySecrets.html on 9/12/12.