Using Speech Patterns to Model the Dimensions of *Teamness* in Human-Agent Teams

Emily Doherty
emily.doherty@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Cara A. Spencer
cara.spencer@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Lucca Eloy
lucca.eloy@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Nitin Kumar
nitin.kumar@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Rachel Dickler
rachel.dickler@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Leanne Hirshfield
leanne.hirshfield@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Teamness is a newly proposed multidimensional construct aimed to characterize teams and their dynamic levels of interdependence over time. Specifically, teamness is deeply rooted in team cognition literature, considering how a team's composition, processes, states, and actions affect collaboration. With this multifaceted construct being recently proposed, there is a call to the research community to investigate, measure, and model dimensions of teamness. In this study, we explored the speech content of 21 human-human-agent teams during a remote collaborative search task. Using self-report surveys of their social and affective states throughout the task, we conducted factor analysis to condense the survey measures into four components closely aligned with the dimensions outlined in the teamness framework: social dynamics and trust, affect, cognitive load, and interpersonal reliance. We then extracted features from teams' speech using Linguistic Inquiry and Word Count (LIWC) and performed Epistemic Network Analyses (ENA) across these four teamwork components as well as team performance. We developed six hypotheses of how we expected specific LIWC features to correlate with self-reported team processes and performance, which we investigated through our ENA analyses. Through quantitative and qualitative analyses of the networks, we explore differences of speech patterns across the four components and relate these findings to the dimensions of teamness. Our results indicate that ENA models based on selected LIWC features were able to capture elements of teamness as well as team performance; this technique therefore shows promise for modeling of these states during CSCW, to ultimately design intelligent systems to promote greater *teamness* using speech-based measures.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; **Visualization techniques**; **Interaction techniques**.

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614121

KEYWORDS

automatic speech recognition; collaboration; teamness; network analysis

ACM Reference Format:
Emily Doherty, Cara A. Spencer, Lucca Eloy, Nitin Kumar, Rachel Dickler, and Leanne Hirshfield. 2023. Using Speech Patterns to Model the Dimensions of *Teamness* in Human-Agent Teams. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614121

1 INTRODUCTION

A challenge that has long faced team science research is how to measure essential, multifaceted teamwork constructs on the team level, as teamwork unfolds over time, rather than simply aggregating individual-level measurements. This challenge has been exacerbated by recent advancements in human-agent teaming (HAT) that have blurred the line between human and agent roles, relationships, and dependencies. As Cooke describes, rather than measuring how individuals in a team function collectively, there is a need for objective, longitudinal measurements that reflect changes in team cognition, composition, and team processes as tasks unfold over time [7].

Cooke's recently dubbed term, teamness, is proposed to be considered within the dimensions of **team composition, role heterogeneity, diversity of shared goals and identity, authority structure, and degrees of interdependence** [7]. While this teamness framework takes a crucial step toward developing a better understanding of today's complex HATs, the teamness authors note that these dimensions require further development, measurement, and testing [7]. **We posit that the dimensions of teamness may be measured, in part, through sub-facets of individual and team-level measures of affective states, degrees of trust, mental workload levels, and team processes, which we describe next.**

1.1 Bridging teamwork measures with teamness

To bridge the teamness dimensions with the aforementioned measurements, we refer to the 'ABCs' of teamwork literature that describes why a team meets their objectives given certain **a**ffective states, **b**ehavioral processes, and **c**ognitive states of team members [3]. The ABCs are a validated framework of measurable teamwork

mechanisms which connect our measurement methods to the teamness dimensions below.

Both individual and team affective states (i.e. valence, and arousal) are affected by affect disposition—making **team composition** important—or behaviors like self- and co-regulation of attitudes, emotion and mood, cooperation, and ingroup **identity** [21]. Affective states are measured using surveys and physiological data, while their associated behaviors can be studied using various methods like speech and facial expression.

Trust is crucial to teamness, especially in HATs, with both affective- and cognitive-based trust playing an important role in **interdependence** via overall trust and reliance [26]. Trusting behaviors, along with a mix of affective and cognitive states, engender team processes like cohesion, communication, and secure team **identity**. Cognitive states also have a reciprocal relationship with trusting behaviors. Behaviors like information sharing and willingness to adapt strategies increase the chances for developing new cognitive states that can positively impact performance [9]. Due to trust's connection with the affective and cognitive states, teams with members exhibiting less trust are more susceptible to team process breakdowns that lead to poor performance [9]. While a lack of trust does not guarantee team performance failure, performance benefits from creativity, cooperation, and coordination are harder to achieve without it.

Connected to teamness through cognitive states and behavior, is mental workload. Cognitive states include the team level knowledge structure and the perception and acquisition of information (e.g, shared mental models) [33]. Within the shared cognitive state could be a **shared goal** which team workload research has shown to decrease mental workload in individual members and team overall [5]. Individual members bring their own measurable cognitive abilities, knowledge, and skills. Typically, more of these attributes improve the team's cognitive state since they contribute to the facilitation of teamwork—an impact of **team composition**. Role differentiation by assigning roles based on the strengths of each member and subsequent role **heterogeneity** is shown to decrease cognitive workload [33]. Agents also contribute to their teammates' cognitive state through their own informational participation, but depending on its characteristics, the deeper mechanisms are harder for humans to intuit so shared states look different.

The foundational definition of team processes comes from Marks et al. [23], who describes these processes as how members work **interdependently** to share resources and organize task work to yield a meaningful outcome [23]. While these processes describe stages of collaboration over time, they may present differently in HATs compared to human teams. Introducing a non-human entity, especially one without clear affective and cognitive processes, to a team has several implications not present in human teams. Human perceptions of an agent (i.e. trust, reliability, fear, suspicion) directly influence team dynamics [2]. Agents typically lack the intelligence, emotion, and other characteristics of their human teammates, which can have negative implications such as lack of trust and higher mental workload [8, 10, 35]. Although currently challenging, agents should be thoughtfully designed to reduce the mental workload of its teammates [42].

Within recent years, there has also been a radical shift in how teams are distributed, causing an increased reliance in virtual communications spanning time and space [27]. Temporally- and spatially-dispersed teams may have different perspectives of **shared goal and identity** that is naturally established in co-located teams [7]. The impact of virtual communication on dispersed team dynamics and communication patterns is not yet fully understood [11]. The level of **interdependence** in virtual teams is also difficult to measure as compared to physical teams, where the execution of sequential and interdependent tasks requires frequent communication [16]. Naturalistic dialogue is especially important in remote teams due to the lack of physical nonverbal indicators (i.e. body language) that contribute to team dynamics [24]. Team cognition has previously been measured dynamically using speech-based measures [20], suggesting speech as a effective measure of teamness.

1.2 Candidate Measures

Surveys and behavioral measures of teamwork can be subjective and obtrusive if they interrupt teams in real-time. With teamwork evolving, it is imperative to take a more naturalistic, multimodal approach to measuring dimensions of teamness using less obtrusive measures that can be applied beyond traditional teams. While non-invasive physiological measurements such as electroencephalogram (EEG), functional near-infrared spectroscopy (fNIRS), galvanic skin response (GSR), eye-tracking, and heart rate variability are telling of one's physical and cognitive states during collaboration, these measurements cannot be applied to artificial agents nor in many real-world environments [28]. However, speech is a rich, multidimensional, team-level metric that can be measured unobtrusively in most team types. Natural dialogue is complex yet informative on its own. In this paper, we analyze speech patterns using word-counting into distinct psychologically meaningful categories [41], described in 3.2. We note that this use of "speech patterns" differs from that in similar multimodal literature that focuses on explicit speech behaviors like question-asking, argument, reasoning, initiation style [29, 32] and prosodic features of speech including pitch and rate [12, 45]. This use of speech patterns, in combination with other multimodal measures, shows great potential to capture the dynamic nature of teamness.

1.2.1 Current study and contributions. In this paper, we aim to evaluate teamness in virtual, human-agent teams using combined survey and speech measures. We posit that four components derived from a combination of affective state, trust, team processes, and workload measures are highly interconnected to Cooke's proposed dimensions of teamness [7]. In this study, we used four components: social dynamics and trust, affect, cognitive load, and interpersonal reliance to split teams into high- or low- component groups for comparison. Speech data was analyzed using Linguistic Inquiry and Word Count (LIWC) to parse out linguistic features of particular interest to teamwork. Epistemic networks were constructed to compare speech patterns of teams for each component. This unique method to visualize speech patterns allowed for the comparison of co-occurrences of LIWC features in high vs. low component teams. Our results suggest that naturalistic speech in teams can be used to model affective state, trust, workload, and team processes;

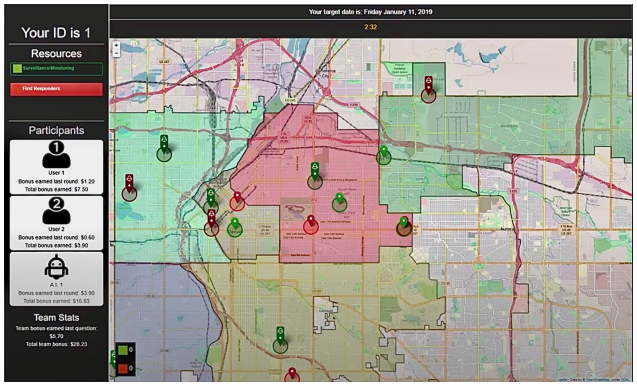

Figure 1: Shared map view of CHART featuring individual scores and a timer reporting the remaining time left in the round.

all of which contribute to measuring the dynamic dimensions of teamness as teams collaborate over time.

2 METHODS

2.1 Participants

42 students, with an average age of 22 years old and 48% female, participated in the study at a large public university. They worked in teams of two human participants and one agent on an experimental task, with a total of 21 teams/sessions. The participants were compensated with a monetary payment of $15/hour, along with a variable cash bonus based on their task score. The recruitment and experimental procedures were approved by the university's Institutional Review Board, and the participants provided informed consent forms before starting the task.

2.2 Experimental Testbed

The experimental task was conducted using the Computer-Human Allocation of Resources Testbed (CHART)[4]. CHART allows teams of two humans and one rule-based agent to collaborate remotely on a spatiotemporal mapping task, where participants are tasked with searching through historical data overlaid on a map to identify trends in unlawful activity. Specifically, given these trends teams are instructed to allocate a limited number of 'crime prevention' resources throughout the city and each crime caught adds to the score. The interface consists of two displays: an interactive map that allows participants to visualize data from specific past dates and categories of offenses, and a shared map where both participants and the agent place their resources, represented as pins. The shared map displays the team's current task score as well as each individual's contribution to the team's score (Fig. 1).

2.3 Survey Measures

Following each round, participants were given a battery of state surveys described next. Within the framework proposed by Marks et al. [23], team processes were measured using items from the established Team Processes survey from Mathieu [25]. We specified that the team consisted of both the other human and the agent. The survey included the items with highest factor loadings and adaptability to our scenario for the processes of coordination, conflict management, goal monitoring, strategy formulation, and cohesion (from "Affect Management" items) with a Cronbach's alpha $\alpha = 0.93$. The participants recorded the extent of their agreement on a Likert scale from Strongly Disagree to Strongly Agree. Participants were also given a visual analog scale to report their emotional valence ("very negative" to "very positive") and arousal ("very sleepy" to "very active"), based on Russell's classic circumplex model of affective states [34]. Three items from the NASA-TLX were presented via an on-screen slider to assess mental demand, temporal demand, and perceived performance [14]. Cognition-based trust, affect-based trust, and teammate-monitoring behavior were measured with the highest factor-loading items from McAllister [26].

2.4 Task Performance

Task score was calculated using the number of events that were 'caught' within the radius of a team's pins. Since real data was used for the task, finding a true solution (optimal pin placements) is computationally intractable for any date. The maximum score obtained by a team was 21 events 'caught' in a single round, and 80 events across all 8 rounds.

2.5 Speech Measurement

Audio of speech was recorded using Zoom, transcribed using Whisper, and processed through the LIWC-22 application. Whisper is an open-source, multi-lingual automatic speech recognition model supporting speech translation and language identification developed by OpenAI [31]. Whisper was chosen given its potential use in future intelligent systems (i.e. conversational agents) to capture user speech and correspondingly apply models to determine interventions. Along with Whisper, stable-ts library was used, which provides timestamp stabilization and thus improves segment-level timestamps [17]. A subset of Whisper-generated transcripts were compared to human-generated transcripts to validate its accuracy. We observed a Word Error Rate (WER) of 2.15%. LIWC-22 computes over 100 features per utterance based on a series of pre-defined dictionaries [6]. These features indicate characteristics of utterances including the number of words spoken and percent of words related to a predefined category. The features from LIWC-22 extracted for the present study are listed in the section 3.2.

3 ANALYSIS

Grounded in the teamness framework, our hypothesis testing pipeline went as follows: derived four teamwork components through survey feature selection, split each component into comparison groups, selected speech features using LIWC characterization, and finally, conducted ENA of those LIWC speech features per each teamwork component.

3.1 Survey Feature Selection

To reduce the number of total survey measures while minimizing information loss, principal component factor analysis (PCA) identified 4 combined measures of valence, arousal, cognitive trust, affective trust, teammate monitoring, mental demand, temporal

demand, perceived performance, and team processes (coordination, conflict management, goal monitoring, strategy, and cohesion). A Kaiser-Meyer-Olkin test deemed the data well suited to factor analysis with a value of 0.85. Subsequent PCA returned 4 components with eigenvalues greater than 1. Varimax rotation with a loading cutoff of 0.3 identified the following composition of factors (Table 1) explaining a cumulative total of 76% variance. We selected these four components as variables of interest relating to several dimensions of teamness.

We define component 1 as **social dynamics and trust**, as it is comprised of every team process measure (loadings 0.86 – 0.92), valence (0.48), cognitive trust (0.63), and affective trust (0.66). This describes the positive social perceptions of one's team, along with their levels of trust. Component 2 represents emotional valence (0.68), arousal (0.76), and perceived performance (0.80). Thus, we label this as **affect**, given its strong link to both affective measures as well as a team's sense of accomplishment. Component 3, comprised of temporal demand (0.87), mental demand (0.68), and arousal (0.37), straightforwardly captures teams' **cognitive workload**. Lastly, cognitive trust (0.47), conflict management (0.36), teammate monitoring (0.87), and inverse mental demand (-0.57) make up component 4. Cognitive trust and monitoring behavior are directly based on one's judgment of their teammate's ability in the task; it follows that as a teammate's competence increases, individuals can rely on them and decrease their own mental demand. Thus, we refer to this component as **interpersonal reliance**. A median split was performed on each component to divide teams into high and low teams per component: Social Dynamics and Trust (median = 0.2265), Affect (median = -0.1409), Cognitive Load (median = -0.0938). and Interpersonal Reliance (median = 0.1437). Along with the survey measures, team success (0,1) was an outcome measure. A median split was used for team score (median = 1.8) to divide teams into high-performing (1) or low-performing teams (0). The median was then subtracted to center the data around 0.

3.2 LIWC Feature Selection

Linguistic content offers insight on a team's processes, affect, and even specific collaborative problem-solving skills [40]. Six LIWC features were selected out of the possible 100 based on prior work using LIWC to model team processes. Specifically, LIWC is commonly used for analyzing speech data from multi-party conversations and prior work has identified several features most reflective of team processes including Analytic Language, Drives Language, Positive Tone, Negative Tone, Cognitive Processes, and Past Tense as described in Table 2.

3.3 Epistemic Network Analysis

Epistemic networks [38] have the potential to meaningfully unpack real-time, conversational speech data during collaborative activities in teams. The epistemic networks are constructed using an optimization routine that accounts for the co-occurrence of features across utterances within conversations. This results in a network with connections between nodes (i.e., LIWC features) weighted to reflect how frequently features co-occur within each conversation. Conversations can be grouped to make comparisons between patterns in discourse associated with particular outcomes (e.g., successful versus unsuccessful task performance). ENA is a valuable modeling approach as it allows for understanding connections between features in discourse, as well as quantitatively and qualitatively comparing patterns in discourse related to specific outcomes [38]. Additionally, the networks that emerge can be used to evaluate whether the model features are able to successfully capture the component of interest and distinguish between outcomes for the construct.

In the present study, the six LIWC features were used as the nodes in the network to compare patterns in discourse in relation to our four team components and team success. Specifically, networks were compared for team conversations according to: High Social Dynamics and Trust versus Low Social Dynamics and Trust, High Affect versus Low Affect, High Cognitive Load versus Low Cognitive Load, and High Interpersonal Reliance versus Low Interpersonal Reliance. All conversations for high versus low groups were determined using a median split as explained in Section 3.1 (resulting in n=65 groups in each network; n = 130 total group conversations). The stanza size for the analysis in all networks was set to a moving window of 4, to best capture patterns occurring within conversations. The networks were compared first quantitatively with a t-test examining differences in the mean centroids of each network. The networks were then compared qualitatively based on the difference in weighted connections between the two networks for each outcome (i.e., the subtracted network).

As a measurement check, we expected to see alignment between participants' self-reported affect and the values of the LIWC features representing affect. We therefore ran a Pearson correlation of the Valence factor from the surveys (ranging from 1= very negative to 5 = very positive) against relevant LIWC features: positive tone and positive emotion. Valence was positively correlated with both positive tone ($r = 0.37$, $p < 0.001$) and positive emotion ($r = 0.34$, $p < 0.001$), confirming that these LIWC features are indeed aligned with participants perceived emotional valence.

4 RESULTS AND INTERPRETATION

An ENA comparison of LIWC features are reported per component, with their relation to the hypotheses specified, and then the same process for team performance. Afterwards, the implications of these results for teamness evaluation are discussed.

4.1 Component 1: Social Dynamics and Trust

We first developed an ENA model to compare speech patterns associated with high vs. low levels of the Social Dynamics and Trust component (Fig. 2). A two sample t-test assuming unequal variance showed that the network for Low Social Dynamics and Trust teams (M = 0.16, SD=0.64, N=65) was statistically significantly different at the α=0.05 level from the network for High Social Dynamics and Trust teams (mean=-0.16, SD=0.28, N=65; $t(87.79)= 3.66$, $p<0.001$, Cohen's d=0.64). Qualitative analysis of the subtracted network revealed that teams with higher social dynamics and trust used significantly greater co-occurrences of analytic, cognitive processes, and past tense language (see Figure 2). This finding suggests that these three LIWC features can explain some of the differences in outcomes regarding how teams dealt with conflict as well as level of trust, validating hypotheses H1, H5, and H6. Notably, the link

Table 1: Components, factors, and loadings yielded by factor analysis

Factor	Component 1: Social Dynamics & Trust	Component 2: Affect	Component 3: Cognitive Load	Component 4: Interpersonal Reliance
Cognitive trust	0.63			0.47
Affective trust	0.66			
Monitoring behavior				0.87
Coordination	0.86			
Conflict management	0.82			0.36
Goal monitoring	0.89			
Strategy formulation	0.86			
Cohesion	0.92			
Valence	0.48	0.68		
Arousal		0.76	0.37	
Mental demand			0.68	-0.51
Temporal demand			0.87	
Perceived performance		0.80		

between cognitive processes and analytical thinking is the most prominent in this network, highlighting the contribution of these features to social dynamics.

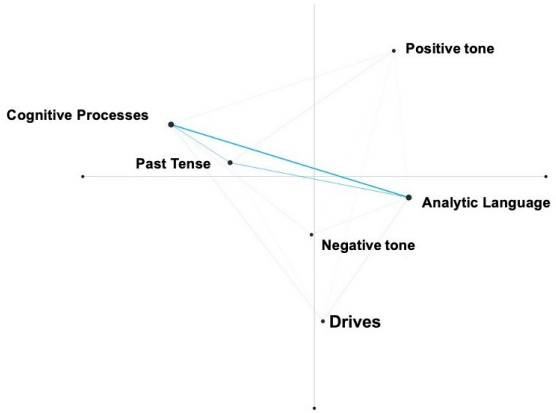

Figure 2: Subtracted ENA network for High Social Dynamics and Trust (blue) – Low Social Dynamics and Trust (red) teams. Blue networks between Cognitive Processes, Past Tense, and Analytic Language translate to greater co-occurences of these features in High Social Dynamics and Trust teams.

4.2 Component 2: Affect

An ENA model to compare speech patterns associated with high vs. low levels of the Affect component was constructed (Fig. 3). A two sample t-test assuming unequal variance showed Low Affect teams (mean=-0.19, SD=0.54, N=65) was statistically significantly different at the α=0.05 level from High Affect teams (mean=0.19, SD=0.29, N=65; t(98.66)= -4.97, p<0.001, Cohen's d=0.87). Groups that rated themselves with greater affect (more positive and energetic) exhibited increased use of analytical thinking, cognitive processes, and past focus language (H1, H5, H6).

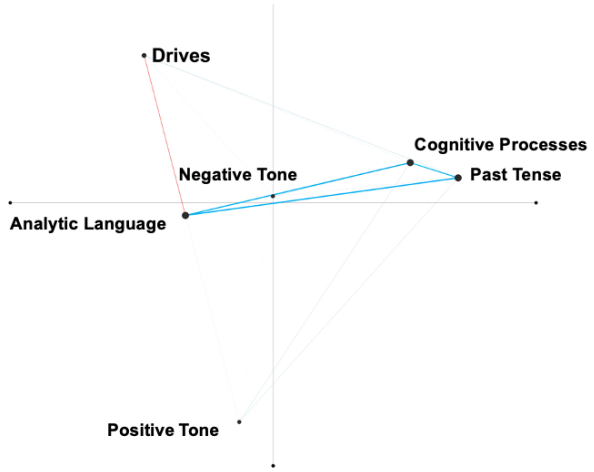

Figure 3: Subtracted ENA network for High Affect (blue) – Low Affect (red) teams. High Affect teams showed greater usage of Cognitive Processes, Past Tense, and Analytic Language, while Low Affect teams used Drives language more often with Analytic language.

4.3 Component 3: Cognitive Workload

An ENA model to compare speech patterns associated with high vs. low levels of the Cognitive Workload component was constructed (Fig. 4). A two sample t-test assuming unequal variance showed Low Cognitive Load teams (mean=-0.15, SD=0.65, N=65) was statistically significantly different at the α=0.05 level from High Cognitive Load teams (mean=0.15, SD=0.42, N=65; t(109.80)= 3.05, p<0.001, Cohen's d=0.54). The analysis revealed that teams with increased cognitive load are most strongly differentiated by increased co-occurrence of cognitive processes language with past focus, analytical thinking, and positive tone language (H1, H3, H5, H6). The strength of the connection between past focus and analytical thinking specifically suggests that teams with higher cognitive workload tended

Table 2: Selected LIWC features and associated hypotheses

LIWC Feature	Description and Hypotheses
Analytic	The "Analytic" feature is a summary variable used to measure logical or abstract thinking (through increased article use) and cognitive complexity (through increased preposition use) [30]. Analytic words have been positively correlated with increased team member effectiveness scores [1], revealing higher levels of interdependence between team members. **H1:** We expect that frequent presence of analytic talk will correspond with higher team processes and with better team performance because logical thinking is necessary to coordinate team activities and successfully complete the CHART task.
Drives	The "Drives" dimension includes words of achievement, affiliation, power, reward, and risk through use of first-person pronouns like "we", "us", and "our" [6]. Drives language is highly correlated with the teamness dimensions of having shared goals and role hierarchy. Drives words have been correlated with the collaborative problem solving facet of maintaining team function [40]. It is important for tasks to simulate real-world risks, as echoed by Cooke [7]. Because CHART mimics real-world risks, we hypothesize that: **H2:** frequent presence of drives will correspond with higher team processes and with better team performance because of the role hierarchy and level of interdependence required of the CHART task.
Positive Tone	Compared to previous versions, LIWC-22 has further classifications of positive and negative emotions into tone categories. These categories now reflect sentiment, rather than emotion by incorporating words related to certain emotions [6]. Assents and positive emotion words measure levels of agreement [41]. It is true that when group members express positive sentiment, it tends to facilitate group functioning [18, 23]. Positive tone words have been positively correlated with higher peer ratings of team effectiveness [1]. Therefore, positive tone language may be associated with higher level of teamness throughout a task. **H3:** We expect that frequent presence of positive tone will correspond with higher team processes and with better team performance as demonstrated in literature.
Negative Tone	Negative affective tone has been associated with poor team performance, decreased group identification [19, 21], and decreased team cooperation [21]. **H4:** We expect that the increased presence of negative tone will correspond with poorer team processes and with lower team performance due to lack of group cohesion.
Cognitive Processes	Cognitive processes words represent causation, discrepancy, differentiation, and insight [6]. This measure can evaluate the degree to which group members engage in reflective thinking. For instance, van Swol et al. (2016) found that groups that had a member with an extreme opinion used less cognitive process language than groups without such members [44]. This may have resulted in a reduced interest in meaningful conversation. Additionally, van Swol et al. (2021) observed that group members who engaged in more perspective-taking utilized more cognitive process language [43]. Both reflective thinking and perspective-taking coincide with the teamness dimensions of role hierarchy and heterogeneity. **H5:** We expect that frequent presence of Cognitive Processes will correspond with higher team processes and with better team performance due to healthy levels of interdependence and heterogeneity of the team (humans and agent).
Past tense	The "Focuspast" feature refers to words spoken in the past tense. Because of the nature of the CHART task, we propose that more successful teams and those with higher team processes with have more frequent use of words in the past tense. **H6:** Because successful task completion requires frequent reference to the historical data, we expect increased past tense language to correspond more frequently with higher team processes and team performance.

to communicate more about past events, perhaps referencing the historical data in the CHART task more frequently. Conversely, teams that maintained a lower cognitive load were more likely to include drives and analytical thinking together in their discussion.

4.4 Component 4: Interpersonal Reliance

An ENA model to compare speech patterns associated with high vs. low levels of the Interpersonal Reliance component was constructed (Fig. 5). A two sample t-test assuming unequal variance showed Low Interpersonal Reliance teams (mean=0.08, SD=0.50, N=65) was statistically significantly different at the α=0.05 level from High Interpersonal Reliance teams (mean=-0.08, SD=0.29, N=65; t(101.85)= 2.36, p=0.02, Cohen's d=0.41). Teams with lower interpersonal reliance displayed higher instances of negative tone (H4) and analytic thinking. This effect seems to follow from instances when individuals could rely less on their teammate and thus had to take on more of the task load themselves. Teams with higher interpersonal reliance had frequent co-occurrences of cognitive process language with drives and analytic language (H1, H2, H5), as well as more use of past tense language (H6).

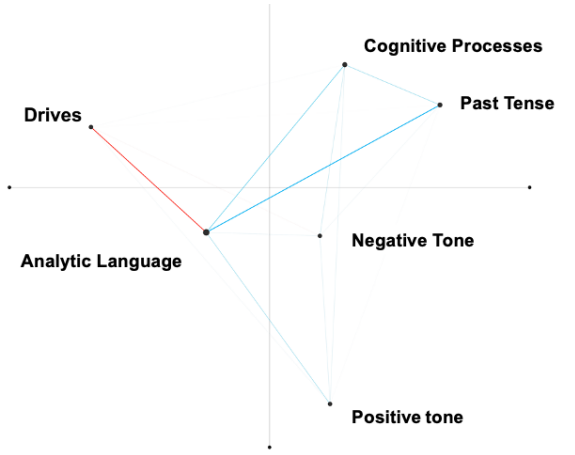

Figure 4: Subtracted ENA network for High Cognitive Load (blue) – Low Cognitive Load (red) teams. Teams with High Cognitive Load more often used Cognitive Processes, Past Tense, and Analytic Language. Teams with Low Cognitive Load frequently used Drives and Analytic Language.

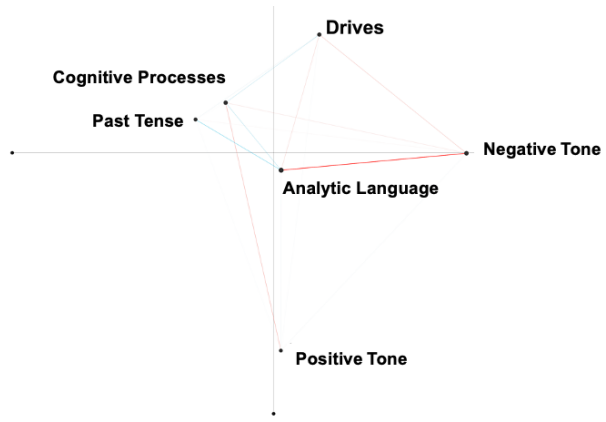

Figure 5: Subtracted ENA network for High Interpersonal Reliance (blue) – Low Interpersonal Reliance (red) teams. Highly reliant teams more often used Negative Tone with Analytic Language, while less reliant teams used Cognitive Processes, Past Tense along with Analytic Language.

4.5 Team Performance

An ENA model to compare speech patterns associated with high vs. low performing teams were constructed using each team's score (Fig. 6). A two sample t-test assuming unequal variance showed Low Performance (mean=0.09, SD=0.45, N=57) was statistically significantly different at the α=0.05 level from High Performance (mean=-0.07, SD=0.30, N=73; t(93.92)= 2.34, p=0.02, Cohen's d=0.43). High performing groups had more co-occurrences of positive tone and cognitive processes, confirming hypotheses H3 and H5. Low performing teams had more instances of past tense, drives, and analytic language, contrary to our hypotheses: H1, H2, and H6. However, low-performing teams used more negative tone, confirming H4.

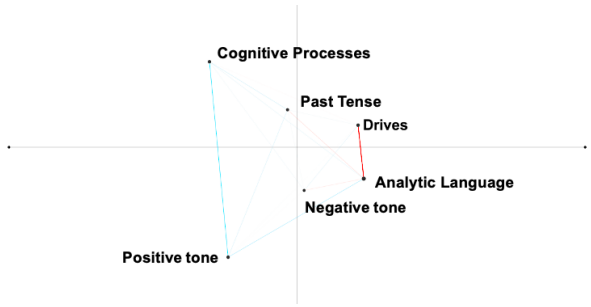

Figure 6: Subtracted ENA network for High Performing (blue) – Low Performing (red) teams. Teams with higher performance used Cognitive Processes language with Positive Tone, while lower performing teams typically used Drives with Analytic Language.

4.6 Interpretations

A few LIWC feature comparison results that stood out for their unexpected relationships or their potential significance in the effort to evaluate teamness. First, increased co-occurrences of drives and analytic language were observed in low affect, low cognitive load, and low performing teams. This suggests that drives language may not be indicative of ideal teamwork language (contrary to H2). First-person plural pronouns (identified as drives language) can indicate increased team cohesion, better performance, and a greater sense of group identity [36, 39, 46] as it decreases **hierarchical** challenges and promotes group communication [46]. However, use of first-person plural pronouns can also indicate use of the Royal We. The Royal We refers to the use of 'we' language by a superior figure to really mean 'you' rather than 'us', signifying role hierarchical issues [41]. This negative relationship between first-person plural language and group cohesiveness has been observed in prior work [13]. The presence of drives language in low workload and low affect groups also support drives as an indicator of poor teamness. This is because teams typically perform better with a healthy amount of cognitive workload and moderate affect.

While drives language may not be the best measurement of optimal teamwork, it was used more frequently among teams with higher inter-reliability. This suggests that teammates relying too heavily on each other may translate to degraded team processes or performance. This aligns with our theory that drives language may be representative of role **hierarchy** and **interdependence** of a team. More drives language will be present in groups that show higher, unhealthy levels of **interdependence**.

Increased co-occurrences of analytic, cognitive processes, and past tense language were observed in high social dynamics and trust, high affect, and high cognitive load groups (confirming H1, H5, and H6). This suggests an noteworthy relationship of these features with higher team processes. We posit that use of cognitive language represents role **hierarchy** and **heterogeneity** dimensions of teamness while analytic language demonstrates **interdependence**. This

implies that teams with higher social dynamics and trust likely engaged in reflective thinking more openly than teams with lower trust. High affect teams, motivated by having a shared goal, used these speech features to maintain team function and role hierarchy. The presence of these features in high cognitive workload groups also demonstrates that use of these linguistic features may help groups manage cognitive load in an effective way. As for past tense language use, future work could investigate if this language is important to quantifying teamness or if it is specifically important to our task.

Regarding significant differences in use of tone, higher performing teams used more positive tone language than lower performing teams (confirming H3), while lower interpersonal reliant groups used more negative tone language than highly reliant teams (rejecting H4). Tone is indicative of team identification, which describes how one self-identifies using the entire team's characteristics and is closely related to performance outcomes [21]. Frequent positive tone leads to higher levels of social integration and subsequently better performance [22], while negative tone is associated with weakened team identification and cooperation [21]. Our results of greater use of positive tone language in higher-performing groups are consistent with previous findings and suggest a greater sense of team identity [22]. Because the CHART task requires high levels of interpersonal reliance through information sharing, greater negative tone usage is consistent with degraded team identification among low interpersonal reliant teams. High performing teams likely used more encouragement and social language, which is consistent with better performance [9]. While groups with less reliance on each other likely used more negative tone words due to their lack of unity. A more balanced approach to reliance is needed in teams when performing a collaborative task.

A few noteworthy interpretations arise from these results that should be considered when designing real-time teamness-rooted interventions:

- Drives language may not be a well-suited measure of positive teamness or high performance. However, it may indicate too much reliance between team members leading to nonideal levels of **interdependence**.
- Cognitive processes and analytic language were observed in higher team processes groups. While we argue these two features' relation to the the teamness dimensions of **role hierarchy, heterogeneity, and interdependence**, this requires further validation in future work.
- Although positive and negative tone can offer some general information about how a team is interacting, tone was not a very informative measure of the dimensions of teamness.

5 CONCLUSIONS

5.1 Study Limitations

While our study provides valuable insights into using speech to model the dimensions of teamness, there are several limitations that must be acknowledged. First, Whisper is a new transcription system and some features may be unstable, impacting some of the LIWC results. Second, while speech is a valuable measure of team communication, our findings could be further strengthened by incorporating additional multimodal measures (i.e. eye gaze, gesture, physiological data) of team interaction. Future research should explore the potential benefits of using multiple measures in combination with speech to better understand the complexities of team collaboration. Third, teamness is a novel construct that requires further development through the identification and measurement of interaction-based dimensions [7]. While our study provides a valuable starting point for this research, further validation is required to capture the full range of teamness dimensions. We encourage future work to expand upon our initial findings between teamness dimensions and associated LIWC features as not all of our hypotheses were correct. Finally, surveys are subjective and obtrusive as they often interrupt simulation of a real-world task. While we took steps to minimize disruption by distributing surveys in between task rounds, future studies should consider alternative methods for collecting data that minimize the impact on team members and do not interfere with team collaboration.

5.2 Conclusions and Future Work

This study marks a significant step toward the quantification of teamness in HATs, with implications for the development of technology to support team processes in real-time. Analysis of epistemic networks comparing speech patterns characterized by LIWC features revealed significant differences between high and low teamwork component groups. When taken in real-time, these speech-based measures can inform the design of intelligent systems that support team processes longitudinally. For example, co-occurrences of analytic and negative tone language indicating low interpersonal reliance could result in a real-time intervention to increase such reliance. A reliance-building intervention may prompt or encourage teammates to increase transparency by communicating their knowledge and reasoning behind an action. The findings suggest that incorporating multimodal data (e.g. physiology, eye-gaze, gesture) into discourse analysis could further improve the accuracy of performance predictions and deepen our understanding of team collaboration dynamics. With recent advances in natural language processing such as the development of conversational agents and proliferation of chatGPT [15, 37], the information we can glean from human speech and the ability to use an agent's speech for more teamwork related functions will continue to grow. The findings suggest that incorporating multimodal data (e.g. physiology, eye-gaze, gesture) into discourse analysis could further improve the accuracy of performance predictions and deepen our understanding of team collaboration dynamics.

ACKNOWLEDGMENTS

This research was supported by the Army Research Office (#W911NF-19-1-0401ARO) and by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of ARO or NSF.

REFERENCES

[1] Rohan Ahuja, Daniyal Khan, Danilo Symonette, Shimei Pan, Simon Stacey, and Don Engel. 2020. Towards the Automatic Assessment of Student Teamwork. In *Companion Proceedings of the 2020 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '20)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/3323994.3369894

[2] Albert Bandura. 1999. Social Cognitive Theory: An Agentic Perspective. *Asian Journal of Social Psychology* 2, 1 (1999), 21–41. https://doi.org/10.1111/1467-839X.00024

[3] Suzanne T Bell, Shanique G Brown, Anthony Colaneri, and Neal Outland. 2018. Team composition and the ABCs of teamwork. *American Psychologist* 73, 4 (2018), 349.

[4] Philip Bobko, Leanne Hirshfield, Lucca Eloy, Cara Spencer, Emily Doherty, Jack Driscoll, and Hannah Obolsky. 2023. Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. *Theoretical Issues in Ergonomics Science* 24, 3 (2023), 310–334.

[5] Clint A Bowers, Curt C Braun, and Ben B Morgan Jr. 1997. *Team workload: Its meaning and measurement.* Psychology Press, 97–120.

[6] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022), 1–47.

[7] Nancy J. Cooke, Myke C. Cohen, Walter C. Fazio, Laura H. Inderberg, Craig J. Johnson, Glenn J. Lematta, Matthew Peel, and Aaron Teo. 2023. From Teams to Teamness: Future Directions in the Science of Team Cognition. *Human Factors* (2023). https://doi.org/10.1177/00187208231162449

[8] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2020. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 131–139.

[9] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of applied psychology* 101, 8 (2016), 1134.

[10] Lucca Eloy, Emily J. Doherty, Cara A. Spencer, Philip Bobko, and Leanne Hirshfield. 2022. Using fNIRS to Identify Transparency- and Reliability-Sensitive Markers of Trust Across Multiple Timescales in Collaborative Human-Human-Agent Triads. *Frontiers in Neuroergonomics* 3 (2022). https://doi.org/10.3389/fnrgo.2022.838625

[11] J. Alberto Espinosa, Ning Nan, and Erran Carmel. 2015. Temporal Distance, Communication Patterns, and Task Performance in Teams. *Journal of Management Information Systems* 32, 1 (2015), 151–191. https://doi.org/10.1080/07421222.2015.1029390

[12] Samantha Finkelstein, Stefan Scherer, Amy Ogan, Louis-Philippe Morency, and Justine Cassell. 2012. Investigating the influence of virtual peers as dialect models on students' prosodic inventory. In *Third Workshop on Child, Computer and Interaction*.

[13] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.

[14] Sandra G Hart and Lowell E Staveland. 1988. *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research.* Vol. 52. Elsevier, 139–183.

[15] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.* https://doi.org/10.48550/arXiv.2301.01768

[16] Guido Hertel, Udo Konradt, and Borris Orlikowski. 2004. Managing distance by interdependence: Goal setting, task interdependence, and team-based rewards in virtual teams. *European Journal of work and organizational psychology* 13, 1 (2004), 1–28.

[17] jianfch. 2023. stable-ts. https://github.com/jianfch/stable-ts

[18] Aimée A Kane and Lyn M van Swol. 2022. Harnessing a language analysis perspective to uncover emergent group processes. (2022).

[19] Thomas Kessler and Susan Hollbach. 2005. Group-based emotions as determinants of ingroup identification. *Journal of Experimental Social Psychology* 41, 6 (2005), 677–685.

[20] Preston A Kiekel, Nancy J Cooke, Peter W Foltz, and Steven M Shope. 2001. Automating measurement of team cognition through analysis of communication data. *Usability evaluation and interface design* (2001), 1382–1386.

[21] Chieh-Peng Lin, Hongwei He, Yehuda Baruch, and Blake E Ashforth. 2017. The effect of team affective tone on team performance: The roles of team identification and team cooperation. *Human Resource Management* 56, 6 (2017), 931–952.

[22] Hector P Madrid and Malcolm Patterson. 2021. Affect and proactivity in teams. In *Emotion and Proactivity at Work*. Bristol University Press, 215–236.

[23] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of management review* 26, 3 (2001), 356–376.

[24] Shannon L. Marlow, Christina N. Lacerenza, and Eduardo Salas. 2017. Communication in virtual teams: a conceptual framework and research agenda. *Human Resource Management Review* 27, 4 (2017), 575–589. https://doi.org/10.1016/j.hrmr.2016.12.005

[25] John E Mathieu, Margaret M Luciano, Lauren D'Innocenzo, Elizabeth A Klock, and Jeffery A LePine. 2020. The development and construct validity of a team processes survey measure. *Organizational Research Methods* 23, 3 (2020), 399–431.

[26] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38, 1 (1995), 24–59.

[27] Sarah Morrison-Smith and Jaime Ruiz. 2020. Challenges and barriers in virtual teams: a literature review. *SN Applied Sciences* 2 (2020), 1–33.

[28] Catherine Neubauer, Kristin E Schaefer, Ashley H Oiknine, Steven Thurman, Benjamin Files, Stephen Gordon, J Cortney Bradford, Derek Spangler, and Gregory Gremillion. 2020. *Multimodal Physiological and Behavioral Measures to Estimate Human States and Decisions for Improved Human Autonomy Teaming.* Report. CCDC Army Research Laboratory.

[29] Bhargavi Paranjape, Zhen Bai, and Justine Cassell. 2018. Predicting the temporal and social dynamics of curiosity in small group learning. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19*. Springer, 420–435.

[30] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver. 2014. When small words foretell academic success: the case of college admissions essays. *PLoS One* 9, 12 (2014), e115844. https://doi.org/10.1371/journal.pone.0115844

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[32] Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In *Ninth International Conference on Spoken Language Processing*.

[33] Yehudit Reuveni and Dana Rachel Vashdi. 2015. Innovation in multidisciplinary teams: The moderating role of transformational leadership in the relationship between professional heterogeneity and shared mental models. *European Journal of Work and Organizational Psychology* 24, 5 (2015), 678–692.

[34] James A Russell and Geraldine Pratt. 1980. A description of the affective quality attributed to environments. *Journal of personality and social psychology* 38, 2 (1980), 311.

[35] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29.

[36] J. B. Sexton and R. L. Helmreich. 2000. Analyzing cockpit communications: the links between language, performance, error, and workload. *Hum Perf Extrem Environ* 5, 1 (2000), 63–8. https://doi.org/10.7771/2327-2937.1007

[37] Abdulla Shafeeg, Ilman Shazhaev, Dimitry Mihaylov, Arbi Tularov, and Islam Shazhaev. 2023. Voice Assistant Integrated with Chat GPT. *Indonesian Journal of Computer Science* 12, 1 (2023).

[38] David Williamson Shaffer, Wesley Collier, and Andrew R Ruis. 2016. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics* 3, 3 (2016), 9–45.

[39] Niklas K Steffens and S Alexander Haslam. 2013. Power through 'us': Leaders' use of we-referencing language predicts election victory. *PloS one* 8, 10 (2013), e77952.

[40] Angela EB Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D Duran, Valerie Shute, and Sidney K D'Mello. 2019. I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.

[41] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[42] Koen van de Merwe, Steven Mallam, and Salman Nazir. 2022. Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors* (2022), 00187208221077804.

[43] Lyn M Van Swol, Paul Hangsan Ahn, Andrew Prahl, and Zhenxing Gong. 2021. Language use in group discourse and its relationship to group processes. *SAGE Open* 11, 1 (2021), 21582440211001852.

[44] Lyn M. Van Swol, Andrew Prahl, Miranda R. Kolb, Emily Acosta Lewis, and Cassandra Carlson. 2016. The language of extremity: The language of extreme members and how the presence of extremity affects group discussion. *Journal of Language and Social Psychology* 35 (2016), 603–627. https://doi.org/10.1177/0261927X16629788

[45] Nigel G Ward. 2019. *Prosodic patterns in English conversation.* Cambridge University Press. https://doi.org/10.1017/9781316848265

[46] Mona Weiss, Michaela Kolbe, Gudela Grote, Donat R. Spahn, and Bastian Grande. 2018. We can do it! Inclusive leader language promotes voice behavior in multiprofessional teams. *The Leadership Quarterly* 29, 3 (2018), 389–402. https://doi.org/10.1016/j.leaqua.2017.09.002