I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving

ANGELA E.B. STEWART, University of Colorado Boulder, USA
HANA VRZAKOVA, University of Colorado Boulder , USA
CHEN SUN, Florida State University , USA
JADE YONEHIRO, University of California, Davis, USA
CATHLYN ADELE STONE, University of Colorado Boulder , USA
NICHOLAS D. DURAN, Arizona State University, USA
VALERIE SHUTE, Florida State University, USA
SIDNEY K. D'MELLO, University of Colorado Boulder, USA

Collaborative problem solving (CPS) is a crucial 21st century skill; however, current technologies fall short of effectively supporting CPS processes, especially for remote, computer-enabled interactions. In order to develop next-generation computer-supported collaborative systems that enhance CPS processes and outcomes by monitoring and responding to the unfolding collaboration, we investigate automated detection of three critical CPS process – construction of shared knowledge, negotiation/coordination, and maintaining team function – derived from a validated CPS framework. Our data consists of 32 triads who were tasked with collaboratively solving a challenging visual computer programming task for 20 minutes using commercial videoconferencing software. We used automatic speech recognition to generate transcripts of 11,163 utterances, which trained humans coded for evidence of the above three CPS processes using a set of behavioral indicators. We aimed to automate the trained human-raters' codes in a team-independent fashion (current study) in order to provide automatic real-time or offline feedback (future work). We used Random Forest classifiers trained on the words themselves (bag of n-grams) or with word categories (e.g., emotions, thinking styles, social constructs) from the Linguistic Inquiry Word Count (LIWC) tool. Despite imperfect automatic speech recognition, the n-gram models achieved AUROC (area under the receiver operating characteristic curve) scores of .85, .77, and .77 for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively; these reflect 70%, 54%, and 54% improvements over chance. The LIWC-category models achieved similar scores of .82, .74, and .73 (64%, 48%, and 46% improvement over chance). Further, the LIWC model-derived scores predicted CPS outcomes more similar to human codes, demonstrating predictive validity. We discuss embedding our models in collaborative interfaces for assessment and dynamic intervention aimed at improving CPS outcomes.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**

KEYWORDS: Collaborative problem solving; collaborative interfaces; language analysis

ACM Reference format:

Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, 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: Human Computer Interaction.* 3, CSCW. (November 2019), 19 pages. https://doi.org/10.1145/3359296

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
Copyright © ACM 2019 2573-0142/2019/November – ART194 $15.00
https://doi.org/10.1145/3359296

1 INTRODUCTION

Collaborative problem solving (CPS) has been noted as a critical 21st century skill for the modern workforce [69]. CPS involves two or more people engaged in a coordinated attempt to share their skills and knowledge, in order to construct and maintain a joint solution to a problem [44,69]. Effective CPS is dependent on a team's ability to establish common ground as it pertains to the problem space [44], jointly develop a solution that accommodates multiple perspectives [44,56], monitor progress toward the goal [15,46], and establish a positive and supportive climate [23,69].

As the workforce becomes more global and distributed, it is increasingly important for people to effectively engage in CPS in computer-supported, remote settings [69]. Unfortunately, current interfaces for remote collaboration do not capture the rich social signals available in face-to-face interaction [1,51]. As anyone who has experienced a Skype or Google Hangouts meeting can attest, video conferencing technologies still have a long way to go. Technical limitations, such as poor camera resolution, occlusions due to fixed camera placement, undersampling, and delayed sound transmissions, dampen communication of basic social signals, like nodding, breathing changes, or gesturing. This decreased information bandwidth can lead to reduced ability to coordinate action, resulting in lower engagement, cohesion, trust, and team performance [51].

To address this limitation, researchers in computer supported collaborative work have attempted to develop interfaces that support remote collaboration by leveraging low-level non-verbal signals. For example, real-time sharing of eye gaze makes each partner aware of the locus of attentional focus of the other [29,49,63]. Indeed, gaze sharing can improve action coordination and mutual understanding (and consequently the collaboration itself) by allowing partners to implicitly reference items without the need of verbal cues [10,50]. Conversely, other systems rely on making an individual aware of their own behavioral patterns, rather than those of teammates. For example, researchers have used eye gaze, head orientation, and interaction information to detect attentional lapses during dyadic online gameplay and text-based conversations, and uses these data to display screen content an individual missed during those lapses [20]. Similarly, speech production has been monitored to warn individuals of interruptions in collaborations [6,14]. Such systems rely on processing basic behavioral cues and providing feedback based on best practices related to these behaviors.

However, it is likely that supporting remote collaboration is much more complex than leveraging low-level signals, like eye gaze or speech production. This is because collaboration involves a dynamic interplay involving multiple complex processes that go beyond low-level signals [11]. Some of these processes include emotion co-regulation [5], establishing common ground [44], building rapport [54], and negotiation [15]. Thus, we think that next-generation computer interfaces must go beyond focusing on the low-level behaviors in which a person is engaging. Such systems should provide insight into higher-level constructs of *how well* a person or a team is collaborating. We take a step in this direction by using spoken language to computationally model the following three complex processes implicated in multiple CPS frameworks [15,23,44]: *construction of shared knowledge*, *negotiation/coordination*, and *maintaining team function*. In doing so, we contribute to our long-term goal of developing intelligent interfaces for collaboration with automated assessment of unfolding CPS processes to provide feedback and/or trigger dynamic interventions aimed at improving collaborative processes and outcomes.

1.1 Background and Related Work

We first discuss theoretical CPS frameworks to ground our three CPS facets in related social science theory. Then, we discuss the state-of-the-art in computer interfaces that support collaboration, as it is our eventual goal to improve these systems with our models. Finally, we discuss computational models of socio-cognitive processes associated with effective CPS, and specifically focus on language-based models most similar to our work.

1.1.1 Theoretical Frameworks of Collaborative Problem Solving

CPS involves a complex set of interlinked behavioral patterns and social signals that dynamically unfold and index high-level collaborative processes, such as negotiation [15], active participation [23], and establishing common ground [23,44]. In the Assessment and Teaching of Twenty-first Century Skills

(ATC21S) framework [18,23], CPS is defined by social and cognitive skills. Social skills focus on managing the team and oneself. There are three facets of social skills necessary for effective collaboration. First, a teammate must participate in the collaboration by being willing and ready to interact with other teammates. Second, a teammate must engage in perspective-taking, where they see the problem from another teammate's viewpoint. Finally, teammates must engage in effective social regulation processes, where they negotiate and compromise as well as harness individual team members' strengths. Cognitive skills focus on managing the task itself (i.e. the problem-solving part of CPS). In order to effectively problem solve, teammates must engage in task regulation where they analyze the problem, make a plan, and move the collaboration forward to achieve a solution. Finally, there should be learning and knowledge building as a result of the collaboration

Similarly, the PISA framework [69] defines three CPS competencies that interact with four problem-solving processes, resulting in 12 levels to characterize CPS skills [17,60]. The first CPS competency involves establishing common ground among team members. Teammates should communicate their knowledge and ideas proactively while working to understand others' ideas, and establishing shared meaning. The second is taking appropriate action where teammates should provide reasons to support their solution proposals and negotiate with others to achieve a consensual solution plan. The third CPS competency involves maintaining a functioning team, which involves each teammate understanding their role in their team, monitoring for communication breakdowns, and adapting when a breakdown occurs. As noted above, these CPS competencies interact with the following four problem-solving processes: (1) a team must explore and understand the problem; (2) then they must organize information and integrate it with personal knowledge, via appropriate representations of the knowledge, and by formulating hypotheses; (3) they must plan and execute their solution, (4) the team must monitor their plan and reflect on how to improve it, as warranted.

Thus, these two frameworks define CPS through actions that move the team towards their problem-solving goal, and create a positive team environment. In order to effectively engage in CPS, individuals in the team must not only complete the task at hand, but they must do so in a way that respects their teammates and maintains a positive team dynamic.

1.1.2 Computer-Supported Collaboration Systems

Researchers in computer-supported cooperative work have developed and evaluated a number of systems that provide collaboration support. These systems assess what behaviors a collaborator is engaging in and provide feedback based on theoretical best practices. Such systems use spoken contributions [6,14], eye gaze [14,48], head orientation [20], and interaction patterns with the computer [13,20], to monitor interruptions [6,14], focus of attention [14,20,48], and ineffective problem-solving strategies [13], which can result in increasing participant awareness of the unfolding collaboration process [14,20,48]. For example, Faucett et al. [14] visualized interruptions, as well as whether or not the face was in screen or the individual was looking at the screen, in remote conversations in order to encourage them to engage in effective behaviors. Rather than simply monitoring behaviors and providing feedback, systems have also been designed to augment human-interaction [20,36]. For example, automated monitoring of language has been used to generate task lists of potential action items for meeting attendees [36].

One exemplary system in the computer-supported collaboration domain was designed and validated by Gutwin et al. [20]. They monitored collaborator attention in virtual interactions and dynamically intervened to mitigate negative effects of brief, but frequent attentional disconnections (e.g., where a user missed important screen content). Attention was monitored using webcam-based face detection to determine if a person was looking at the screen, the order of windows to determine whether the relevant window was in focus, and mouse clicks to monitor if the user was clicking outside the relevant window. They used three "catch-up" techniques to visualize changes in a shared virtual interface. First, they highlighted objects that changed during attentional disconnections to increase the visual saliency of these objects. Second, they used visual traces to show how objects moved across the screen or if objects were removed. Finally, they implemented a replay mechanism that displayed a video of missed screen content at

an increased speed. They validated their system on dyads engaged in text chat or gaming applications. They found that users preferred the three catch-up visualization techniques to no visualization. Additionally, the optimal catch-up technique depended on the application (e.g., highlighting was sufficient for simple visual interfaces, replay was not useful for games with fine-grained time constraints). Thus, a variety of work has created systems that leverage low-level behaviors, with the goal of encouraging effective collaborations.

1.1.3 Computational Models of Socio-cognitive Processes

Prior research has used behavioral patterns to model CPS and related phenomena. For example, research has identified patterns of facial expressions, eye gaze, head movements, physiology, speech, and interface interaction that are associated with teammates' active participation [2], effective communication [26,37,64], collective intelligence [7], role in the team [16], agreeableness [33], and task performance [61]. Social-signal processing and multimodal modeling have also been instrumental in understanding conversational characteristics, such as turn-taking [30], speaker's influence [40], or conversational ice-breaking [27]. Hung and Gatica-Perez [24], in particular, linked team cohesion – a sense of belonging to the team – to the audio-visual channels of task-oriented (as opposed to socially-oriented) groups . Related to team cohesion, rapport pertains to interpersonal relationships between group members developed over time [54]. Müller et al. [38] detected rapport loss during open-ended conversation in small groups using an ensemble of support vector machines trained on facial expressions. Empathy, the ability to share the feelings of others [12,59], is another complex socio-cognitive construct important in collaborative interactions [31]. Ishii et al. [28] analyzed how verbal features and gaze transition patterns during turn-changing and turn-keeping events could predict high- and low- empathy levels during multiparty conversations. Thus, prior work has demonstrated the feasibility of detecting socio-cognitive processes from behavioral signals, though it has not directly focused on collaborative problem solving.

1.1.4 Language Based Models of Collaborative Problem Solving

Our work is grounded in the research on linguistic modeling of CPS processes and outcomes, such as idea sharing [15,22], argumentation [45], and task performance [39]. Researchers typically index the content of the collaboration by quantifying the frequency of words and word phrases (n-grams) [15,22,39,43,45], part-of-speech tags [39,45], rare words [45], emoticons [22], and punctuation [15,22,45]. Research has further employed language categorization techniques, such as sentiment analysis to indicate positive or negative attitude [39], cohesion to measure similarity of sentences [39], and dialogue act tagging to indicate the function of an utterance [28].

Most similar to the current study is work by Hao et al. [22], where the researchers used linear chain conditional random fields on sequential text chats between dyads to detect four important CPS facets: *sharing ideas*, *negotiating ideas*, *regulating problem-solving activities*, and *maintaining communication*. They pre-selected unigrams (words), bigrams (two-word phrases), and emotional text symbols based on hypothesized relationships to the CPS facets. They found that sequential modeling achieved an average accuracy of 73.2% and outperformed standard non-sequential classifiers and a baseline accuracy of 29%. Flor et al. [33] individually modeled 31 behavioral indicators (e.g., providing task-relevant information is an indicator of sharing ideas; asking a teammate to clarify is an indicator of negotiating ideas) of the same four CPS facets. In addition to n-gram and punctuation frequencies, they automatically tagged text chats with dialog acts. They trained a classifier on dialog acts from an unrelated dataset and applied this model to their dataset to produce probabilistic dialog act features. Using the n-grams and automatically tagged dialog act features achieved an accuracy of 60.3% with a Naïve Bayes classifier, which beats their majority class baseline of 24.9%. Thus, these studies point to the potential of using written text to monitor CPS.

1.2 Current Study and Novelty

We use spoken language to automatically model the CPS processes (or facets) of construction of shared knowledge, negotiation/coordination, and maintaining team function while 32 triads used video conferencing software to complete a challenging 20-minute computer programming task. Our study is novel compared to previous approaches at this problem in several respects. First, we directly model three complex facets derived from a theoretical framework of CPS, rather than behaviors related to, but not

exclusive to, CPS (e.g., turn taking; behavioral synchrony). There are a number of systems [6,13,14,20,36,47] that support collaboration by measuring such- behaviors (e.g., speech production, eye gaze) and providing users with appropriate feedback (e.g., speak more, look here). Whereas these systems encourage behaviors that set the stage for positive social interactions, these systems do not model CPS itself, which is a very complex problem as it involves multiple individuals with different personal experiences and backgrounds collaborating to solve a challenging problem.

Second, prior research that has modeled CPS phenomena in remote computer-supported interaction has restricted communication to text chats among dyads [15,22], a somewhat limited communicative medium. In our study, triads remotely collaborated on an open-ended task using video conferencing with screen sharing. Participants could freely communicate via speech, gesture, facial expression, and paralinguistics, yielding a rich set of authentic social behaviors. Further, in this initial work, we use linguistic features to model CPS facets, and it is unknown whether language itself is sufficient for this task given the rich communication environment. Our use of audio also presents novel challenges in language modeling because automated transcription and speech segmentation are imperfect, whereas text chats provide a precise representation of the communication between partners. Thus our work provides an initial exploration into fully automated spoken language-based modeling of CPS facets.

Third, because collecting multimodal data during open-ended multiparty interactions is challenging, most research has used a small number of participants and teams (e.g., 16 participants in one case and only one team in another [28,65]). In contrast, our dataset, consisting of 111 participants across 32 teams, is comparatively larger than most previous work. This sample size affords the opportunity to build team-generalizable models, which was not possible with some prior studies (e.g., [24,27]). Relatedly, whereas the most similar studies in this area [15,22] have focused on dyads, the analysis of triads is also novel and adds to the complexity since it affords seven units of analysis (three individuals, three dyads, and the triad itself) compared to the three units in a dyad (two individuals and the dyad)

Finally, in addition to reporting accuracy of the models, we also demonstrate evidence of their predictive validity by showing that our model estimates of CPS facets predict important collaboration outcomes, with results being comparable to human coded CPS facets. To our knowledge, this has not been done in comparable studies (e.g., [15,22]). Taken together, our work provides critical first step in modeling CPS facets from spoken language during remote collaborations with video conferencing.

2 DATA SOURCE

The dataset we use was collected as part of a larger project [55], but the analyses reported here has not been previously published.

2.1 Participants

Participants were 111 (63.1% female, average age = 19.4 years) undergraduate students from a medium-sized private university who self-reported not having any previous experience with computer programming. Participants were 74.8% Caucasian, 9.9% Hispanic/Latino, 8.1% Asian, 2.7% Other, 0.9% Black, 0.9% American Indian/Native Alaskan, and 2.7% did not report ethnicity. Participants were scheduled as teams of three (37 triads in total) depending on scheduling constraints and 19 participants from ten teams (27% of all teams) indicated they knew at least one person from their team.

2.2 Procedure

Participants were randomly assigned to one of three separate laboratory rooms, equipped with a computer, webcam, and microphone. Using Zoom's video-conferencing and screen sharing functionalities (https://zoom.us), participants could see and hear each other while collaborating. Each participant's audio was recorded on separate streams and screen content was also recorded as a video. One participant (designated participant A) was randomly assigned to interact with the collaborative interface and shared his/her screen content with the others (designated B and C) during the entire study.

The task was to collaboratively develop computer programs using the Minecraft-themed Hour of Code environment (Fig. 1) [66], where each chunk of code is represented as a syntactically-correct interlocking block. The main goal was to assemble blocks to satisfy specific design constraints (see below). Teams first collaboratively completed five introductory lessons and watched three accompanying videos on basic programming principles along with instructions on how to use the environment for about 20-minutes. The purpose of these introductory lessons was to familiarize participants with the coding environment and their teammates. After completing the introductory lessons, or the 20-minutes expired, participants then individually (i.e., screen sharing was disabled) rated their level of satisfaction with their team's performance, communication, cooperation, and agreeableness using a 1- 6 scale.

Next, teams were asked to solve a challenging CPS task where they had 20 minutes to satisfy five constraints: 1) build a 4×4 brick building; 2) use at least one if statement; 3) use at least one repeat loop; 4) build at least three bricks over water; and 5) use 15 blocks of code or less. After completing the challenging CPS task, participants individually completed the same subjective assessments to indicate their perception of the collaboration. Finally, participants individually completed a ten-item researcher-created multiple-choice test to assess their learning of the coding concepts.

2.3 Data Treatment

2.3.1 Data Exclusion

Due to equipment failure, data from four teams were removed because at least one participant in the team did not have an audio file and one was removed because they were missing a screen recording. In total, 32 teams and 96 participants were used in the present analyses.

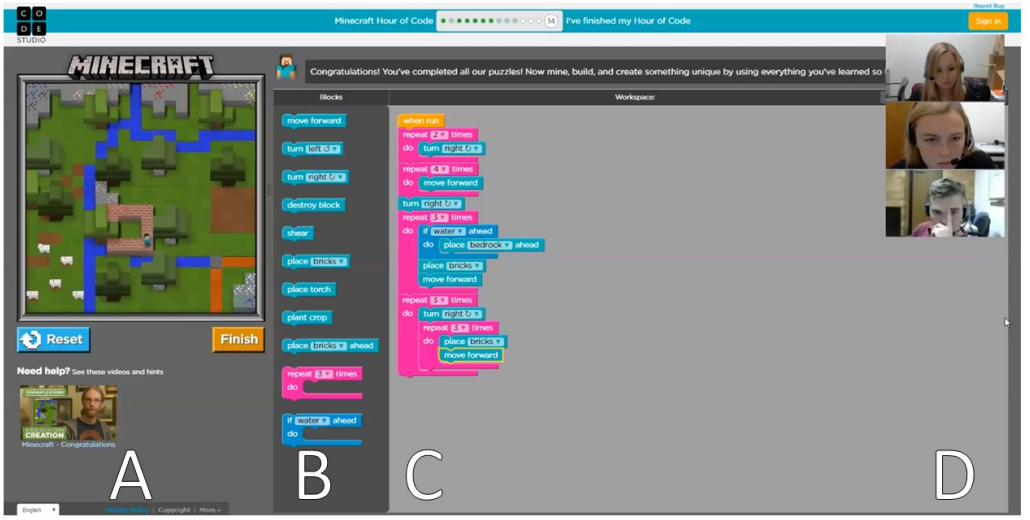

Fig. 1. Collaborative user interface of Minecraft-themed Hour of Code. Participants (D) assemble a chain of code blocks (B) from the code bank (C) and run the animation (A).

2.3.2 Measures

We computed the following measures pertaining to the challenging problem solving task . First, we used each individual's perceptions of their team's performance without any modification. However, we averaged self-reports of communication, cooperation, and agreeableness, also collected after the challenging CPS task, because they were highly correlated (Cronbach's alpha = .89). We scored the 10-time posttest on a scale from 0% to 100%, as an objective measure of learning. The second objective measure involved the task score (i.e., how did each team do on the challenge). For this, two independent raters scored each team's final solution based on the five challenge criteria. The two raters reconciled any

disagreements via discussion. Each criterion was worth a single point, so scores ranged from zero to five with a mean of 2.86 (*SD* = 1.06).

Data was collected over two semesters, with a minor change between semesters, where we added a five-minute warning before the end of the CPS task. Thus, we z-scored group task score, individual posttest score, perception of performance, and perception of collaboration by semester. Additionally, we z-scored self-reported ACT/SAT scores, which were collected as part of a demographics questionnaire.

2.4 Automated Speech Transcription

Each participant's audio files were automatically transcribed using the IBM Watson Speech to Text service [67]. The service generates a transcript with start and stop times for each utterance spoken by each participant. We interleaved transcripts from each participant to produce a one team-level transcript of the collaborative session. Some utterances were incorrectly split into multiple segments. To remedy this, sequential utterances were combined into a single utterance: if (1) they belonged to the same speaker, and (2) there was less than two seconds between the end of one utterance and the start of the next. We decided on this heuristic after assessing the accuracy of utterance segmentation thresholds of 1s, 1.5s, 2s, and 3s. In total, there were 11,163 utterances across the 32 teams for the 20-minute challenging task.

To assess accuracy of the automatic transcription, we manually transcribed a random sample of 10% of the utterances sampled from all participants (total of 1,114 utterances). We computed word error rate (WER) [25], as: (substitutions + insertions + deletions) / (words in human transcript), and set it to zero if the automated transcription indicated speech when there was none (6% of the utterances). The average WER was 45%, indicating considerable imperfections in the transcription. We study the effect of these transcription errors in Section 4.5.

2.5 CPS Framework and Coding of Utterances

We annotated teammates' utterances using a theoretically-grounded and empirically-validated CPS framework [57]. The framework defines three CPS facets: 1) construction of shared knowledge, 2) negotiation/coordination, and 3) maintaining team function. Each facet has three verbal indicators, which form the basis of the coding scheme. Construction of shared knowledge involves sharing ideas and expertise with other teammates and establishing shared understanding among the team. The three indicators for this facet included an individual proposing specific solutions, talking about the givens and constraints of the task, and confirming understanding by asking questions/paraphrasing. Negotiation/coordination is an iterative process to develop and execute a group solution and then revise such solutions as necessary. Its behavioral indicators include providing reasons to support a potential solution, responding to others' questions/ideas, and talking about results. Maintaining team function reflects a positive group dynamic where members are conscious about being part of a team and proactively contribute to its success. Verbal indicators include asking if others have suggestions, complimenting or encourages others, and giving instructions. In total, there were nine verbal indicators (three per facet). Example indicators for a contiguous chunk of one team's transcript are shown in Table 1.

Two human coders, familiar with the coding scheme, were trained to code the utterances for the presence of each indicator. Coders watched video recordings alongside the transcripts and counted the number of times each indicator occurred in an utterance. Coders reached an agreement of .98 (Gwet's AC1) on two five-minute video samples consisting of 254 utterances. The 32 videos were then randomly assigned to the coders, who individually coded their videos.

The resultant indicator counts were binarized (per facet) at the utterance level (since a majority were either 0 or 1). We then applied a logical OR per facet such that if any of the indicators for an utterance was 1, this was taken as positive evidence of the facet for that utterance. In total, 33% of the utterances exhibited evidence of constructing shared knowledge, 15% for negotiation/coordination, and 10% for maintaining team function.

Table 1. Example continuous dialog (human and automatic transcriptions) from a group interaction with the coded indicators and CPS facets. Speaker A is controlling the interaction with the Hour of Code environment.

Speaker	Human Transcription	Automated Transcription	Coded Indicators
A	*Do you want to see if this works at all? Then we can figure out something if it doesn't.*	*do you want to see if this works at all then we can figure out something else it doesn't*	Asks for suggestions (Maintain)
C	*Yeah.*	*if*	Responds to questions/ideas (Neg./Coord.)
B	*We could uh…yeah, just try it. What we do in the beginning is we can like, we can turn right and then place the water blocks and he can turn around again and then just do what we were doing earlier.*	*we can %HESITATION yeah I just try what we do in the beginning as we can light make him turn right place the fix that water block and turn around again and just do what we're doing earlier*	Proposes specific solutions (Constr.)
A	*We can just add another one.*	*we can add another one*	
A	*Right.*	*right*	
C	*Like right there?*	*like right there*	Asking questions/ Paraphrasing (Constr.)
A	*Uh-huh.*	*who*	
B	*Yeah, cause he's standing on that thing, and he comes down and he can't do anything from there.*	*yeah I can see standing and I think it sums down and he can't do anything from there*	Provides reasons (Neg./Coord.); talks about results (Neg./Coord.)
C	*Yeah.*	*yeah*	
A	*Uh-huh.*	*he*	
B	*So maybe in the beginning we can have him fill in that water block, so we don't have, we don't need that water statement then?*	*so maybe in the beginning we can have them fill in that water block so we don't and we don't need that that water statement and*	Proposes specific solutions (Constr.); provides reasons (Neg./Coord.)
A	*So here?*	*so here*	Asking questions/ paraphrasing (Constr.)
B	*Yes. Instead of turning right, he can just turn…see where he's standing.*	*yes it instead of turning right you can just turn all CC standing*	Responds to questions (Neg./Coord.)/ideas; gives instructions (Maintain)

Note: Constr. = Construction of Shared Knowledge; Neg./Coord. = Negotiation/Coordination; Maintain. = Maintaining Team Function

3 SUPERVISED MACHINE LEARNING

Our goal was to automatically model the trained human-raters' codes of construction of shared knowledge, negotiation/coordination, and maintaining team function, in a way that generalizes to new teams. By automatic modeling, we mean that we train supervised classifiers to learn how to generate the human codes from the transcripts without overfitting (i.e., so it the models are applicable to new teams with the same setup). We used language-based features because speech was the primary communication modality, thus language was expected to best reflect the content of the collaboration. Although other multimodal features (e.g., facial expressions, eye gaze) capture aspects of collaboration, they do not index the content of what actually was said. Further, our CPS framework is largely based on verbal indicators,

making language the ideal choice for developing initial modes, which can be subsequently extended with multimodal data.

3.1 Feature Engineering

We derived features using a bag-of-n-grams approach following an open-vocabulary method [52] where counts of words and phrases serve as features. First, utterances were tokenized into individual words using the nltk [4] tokenizer. We experimented with whether to perform word stemming, where word variants are reduced to common roots, using the nltk implementation of the Snowball Stemmer [42]. Similarly, we experimented with removal of stop words (i.e., commonly used words like "a" and "the"), using the Glasgow Information Retrieval Group stop word list [35]. We found that neither word stemming nor stop word removal improved results, so we did not do either in our final models.

One potential downside of using n-grams is that the models might be very specific to our particular domain, thereby reducing generalizability. Thus, we investigated an alternate word encoding method, which used features that would theoretically generalize to other domains. Accordingly, we used the Linguistic Inquiry Word Count (LIWC) [58], which provides occurrences counts of predefined word categories (e.g., affective terms, future terms) obtained from theoretically-grounded and psychometrically-validated dictionaries, to count the proportion of words in an utterance that belong to each of 73 pre-defined LIWC categories. These counts were converted to binary values of whether or not the category was present in the utterance. Any non-zero LIWC categories (i.e., at least one word in the utterance was in that category) were added as a feature. An example automated transcription and corresponding LIWC coding is shown in Table 2. Note that the domain-specific word "turns" is now replaced with the more general word categories "verb", "present focus," "relativity," and "motion".

Table 2. Example coding of LIWC categories based on automatically generated transcriptions.

Automated Transcription	LIWC Transcription
"however trying to build a house like right along right along that water there"	function word - pronoun - impersonal pronoun - article - preposition - adverb - conjunction verb – comparative – cognitive process - tentativeness - differentiation – biological process - ingesting – drives and needs - achievement - relativity - space - home
"I think it's here turns right"	function word - pronoun - personal pronoun - first person singular - impersonal pronoun -auxiliary verbs - adverb - verb - cognitive processes - insight words - present focus - relativity - motion - space
"other repeat four times I think"	function word - pronoun - personal pronoun – first person singular - impersonal pronoun - verb - number – cognitive processes - insight words – differentiation - present focus - relativity - time

3.2 Supervised Classification and Cross Validation

We experimented with the sci-kit learn [68] implementations of the Random Forest and Naïve Bayes classifier, but the former consistently resulted in higher accuracy, so we focus on it.

We used *team-level* 10-fold nested-cross validation. By team-level, we mean that all the utterances for a given team were in the training set or testing set, but never both, which is important for generalizability to new teams. Within each testing fold, the training set was again split into five validation folds for hyperparameter tuning. For each of the five validation folds, a model was fit and scored using every combination of hyperparameters (Section 3.3) via a grid search. Models were scored using area under the receiver operating characteristic curve (AUROC), which assesses the true positive and false positive tradeoff across classification thresholds [21]. Scores for each parameter combination across the five

validation folds were averaged. The hyperparameters which resulted in the highest average AUROC were preserved. A model was then fit on the full training set using these best hyperparameters and predictions were generated on the test fold. These predictions were then pooled over the ten test folds, upon which accuracy metrics were computed.

3.3 Hyperparameter Tuning

We tuned four hyperparameters using the nested cross-validation scheme described above.

N-gram range: We varied the range of n-grams to include unigrams (words) or bigrams (two-word phrases). We chose not to test beyond bigrams because all unique trigrams (and beyond) occurred in less than 1% of the utterances. We did not consider bigrams for models with LIWC features as they are not theoretically meaningful when considering LIWC categories (LIWC does not encode order of the word categories).

Pointwise mutual information: Bigrams were filtered using pointwise mutual information (PMI) [8,34] to ensure that meaningful bigrams (such as "repeat loop") were preserved rather than bigrams that were merely the result of frequent words occurring next to one another (such as "next the"). We tested a low PMI of 2 and a high PMI of 4.

Minimum document frequency: We excluded n-grams that occurred in less than 0%, 1%, or 2% of the training utterances with the specific percentage included as a hyperparameter.

Data balancing method: We tested methods for balancing the training set to address class imbalance. We tested random undersampling, random oversampling, and synthetic minority oversampling technique, using implementations from the imbalanced-learn library [32]. Class distributions for the validation and testing sets were left unchanged. A graphical representation of our modeling pipeline is shown in Fig. 2.

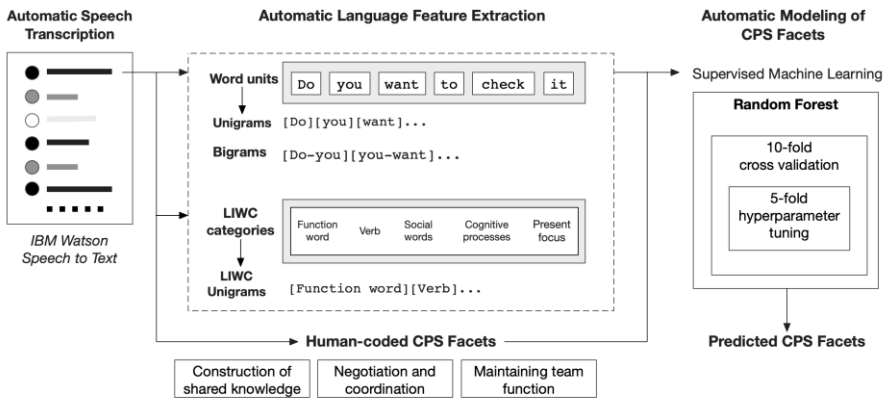

Fig. 2. A graphical representation of our data processing and modeling pipeline. Automatically generated utterances are coded for LIWC features, unigrams, and bigrams. Trained humans also code each utterance for evidence of our three CPS facets. This data is then fed into our machine-learning pipeline.

4 RESULTS

The Random Forest model outputs a likelihood from 0 to 1 that the instance exhibits evidence of the CPS facet (a binary one label). Rather than selecting a probability threshold for binary classification of each instance, we used the area under the receiver operating characteristic curve (AUROC) as our accuracy metric. AUROC assesses the true positive and false positive tradeoff across all prediction thresholds [21] with an AUROC of 0.5 reflecting chance performance.

4.1 Model Accuracies

AUROC values are shown in Table 3 and the correspond ROC curves in Fig. 3. All models performed better than chance, with the n-gram models performing slightly better than the LIWC-category models. Specifically, the n-gram models performed 70%, 54%, and 54% greater than chance for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively. The LIWC-category models performed 64%, 48%, and 46% better than chance for the same facets. The best results were achieved for construction of shared knowledge, ostensibly because of a more balanced training set (i.e., 33% compared to 15% and 10% for the others) and more positive instances to learn from.

Table 3. AUROC values for n-gram and LIWC-category models are shown.

	N-gram	LIWC-category
Construction of Shared Knowledge	.85	.82
Negotiation/Coordination	.77	.74
Maintaining Team Function	.77	.73
Chance	.50	.50

As a follow-up analysis, we combined the models in two ways. First, we averaged the prediction probabilities for the two models, but there was no improvement over the n-gram model (average AUROC was .80 for both models). Second, we combined the automated transcript and LIWC unigrams at the utterance-level and included them in the same model. This joint model also did not yield any improvement over the n-gram model (average AUROC of .79 for combined model versus .80 for n-gram model), so we do not examine it further.

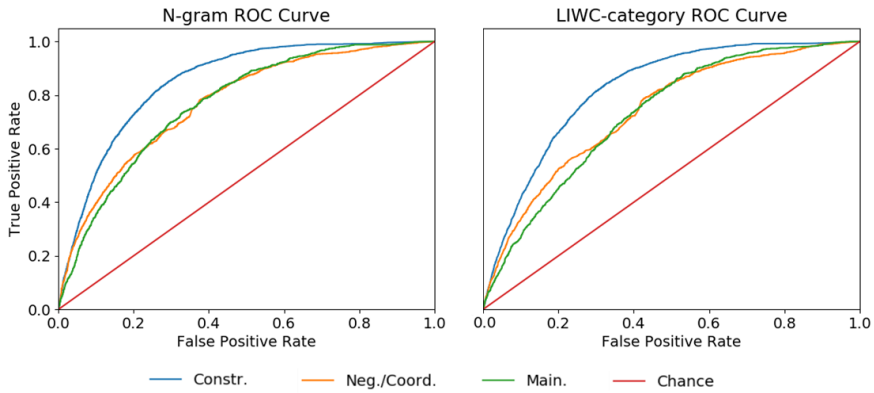

Fig. 3. Receiver operating characteristic curves for the n-gram and LIWC models for all three CPS facets. Note, Constr. = Construction of Shared Knowledge; Neg./Coord. = Negotiation/Coordination; Main. = Maintaining Team Function

4.2 Effect of Team Makeup

We computed three measures of team makeup, in order to investigate whether these background measures might improve the models' accuracy. First, we computed a binary gender makeup variable, setting its value to 1 if all teammates reported the same gender, 0 otherwise. Similarly, we computed a binary racial makeup variable that was 1 if everyone in the team self-reported being the same race. We also computed a binary familiarity variable that was set to 1 if anyone in the team reported knowing anyone else in the team. We modeled the three CPS facets using the same procedure described in Section

3, but we included the three team makeup scores as additional features. These models did not improve prediction accuracy above the n-gram model (mean AUROC values of .78 versus .80 with and without these additional features included, respectively), so we did not include team makeup in our final models.

4.3 Predictive Features.

We identified the top ten n-grams and LIWC categories most strongly correlated with each CPS facet to provide insight into the models. Because a separate model was trained in each of the ten folds, we averaged correlations across all folds after setting the correlation for an n-gram/LIWC category in a fold to zero if it was not included in that fold. Table 4 shows the most correlated n-grams and LIWC categories for each model.

Table 4. Top 10 n-grams and LIWC categories most strongly correlated with each facet.

Model	Facet	Top 10 N-grams
N-gram	Construction of Shared Knowledge	to, the, we, like, forward, and, *yeah*, place, move, water
	Negotiation/Coordination	because, the, I, he, water, think, we, that, it, yeah
	Maintaining Team Function	move, forward, then, turn, and, that, place, repeat, good, times
LIWC-category	Construction of Shared Knowledge	prepositions, space words, relativity words, time words, numbers, motion words, articles, comparisons, discrepancies, auxiliary verbs
	Negotiation/Coordination	causation, negations, insights, auxiliary verbs, articles, ingestion words, first person singular words, differentiation, personal pronouns, male references
	Maintaining Team Function	motion words, relativity, time words, space words, present focus, verbs, future focus, reward words, prepositions, drives

Note: negative correlations are *italicized*.

Recall that the assigned task was to build a 4×4 brick building with multiple constraints, including at least three bricks over water (see Section 2.2). Accordingly, we found that constructing shared knowledge was most strongly predicted by domain-specific n-grams (i.e. "forward", "place", "move", "water"), indicating discussion of possible solutions to the task. Some of the LIWC features predictive of construction of shared knowledge were also domain-specific, such as "numbers," which are presumably for counting the lines of code in their solutions since they were constrained to 15 code lines. However, there were also some domain generalizable features. For example, sharing understanding of problems/solutions (a key indicator, Section 2.5) could be reflected in "comparison" words when teams examine multiple solutions or describe a solution (e.g., "do this *after* that"). Teams could also use first person plural words (e.g. "we", "us") to establish common ground by discussing collective actions the team could take.

Negotiation/coordination was most strongly predicted by n-grams indicative of providing reasons for or against implementing a solution (i.e. "*because* I'm not sure if you're allowed to do that") or discussion related to solution execution and the unfolding on-screen action ("so there's *water* there"). The LIWC features reveal a similar pattern, as reasoning-related categories (e.g., "causation") predicted this facet. Further, teams can discuss results and describe what did/did not occur through "differentiation" and "negation". For example, the utterance "maybe we should have them lay the bricks afterwards because it looks like he's on a different level here and he doesn't recognize that there's water there," describes a reason for implementing a solution based on a result the team just observed. The LIWC categories causation, differentiation, and negation are exemplified by the words "because," "different," and "doesn't," respectively.

Similar to construction of shared knowledge, maintaining team function was correlated with domain specific words, largely due to the high occurrence of those words in general. However, unique words indicating compliments (e.g., "*good*") help discriminate it from construction of shared knowledge. LIWC-categorized "reward words", similarly index compliments and support of teammates.

4.4 Predictive Validity

To what extent do the automated scores predict CPS outcomes? We first correlated (Spearman's rho) task score with team-level scores of the three CPS facets (Table 5), computed by averaging the human-coded utterance scores and the model-predicted utterance-level probabilities for each team. None of the correlations were significant, ostensibly due to the sample size of 32 teams. Instead, we focus on the magnitude of the correlations, with a correlation of 0.2 reflecting a small to medium effect size (Cohen's d of 0.4 [9]). We note that model predicted scores of construction of shared knowledge positively correlated with task scores, whereas human scores did not. In contrast, human scores correlated with negotiation/coordination, but model scores did not. Both human and model scores predicted maintaining team function in the expected positive direction.

Table 5. Spearman correlation between the CPS facets and team task score.

	Construction of Shared Knowledge	**Negotiation/ Coordination**	**Maintaining Team Function**
N-Gram	.21	-.04	.12
LIWC-category	.26	.09	.21
Human	-.09	.21	.15

Next, we used linear mixed effects models [3] to investigate the relationship between the three CPS facets and the following CPS outcome variables assessed at the individual level: posttest score, subjective perception of the team's performance, and of the collaboration process (see Section 2.3.2 for details). This is the recommended analytical approach due to the nested structure of the data where individual participants are nested within teams [41].

We averaged the human-coded utterance scores and the model-predicted utterance-level probabilities for each participant for inclusion as predictors. To examine whether the human-coded and model-predicted scores yielded similar effects, we constructed separate models for each, resulting in 27 models (3 facets × 3 outcome variables × 3 sources [human vs. n-gram vs LIWC-category]). Specifically, each model regressed an outcome variable on one of the facets coded by one of the sources with team identity included as an intercept-only random effect (which adjusts the model intercept per team). We also included each individual's total words spoken, ACT/SAT score, whether the individual knew his/her teammates, and whether the individual was assigned to interact with the environment as control variables as covariates. Results are shown in Table 6. Using the lme4 [3] package in R for the requisite computation, we found that compared to n-gram scores, LIWC scores yielded more similar coefficients to human-coded scores, Specifically, both LIWC model-derived and human-coded scores of construction of shared knowledge positively predicted posttest score. We also found that the human scores of maintaining team function predicted subjective perceptions of the collaboration, but the models' scores did not. Neither the human nor the models' scores of maintaining team function were predictive of subjective perceptions of team performance, though this nonsignificant trend was consistently in the positive direction for both human-coded and LIWC scores. Taken together, both the human- and computer-estimates were more effective at predicting the objective outcomes compared to the subjective ones, ostensibly because the former were based on more objective criterion compared to the latter.

Table 6. Unstandardized regression coefficients for predicting outcome measures from n-gram, LIWC-category, and human-coded scores of CPS facets.

		Dependent Variable		
CPS Facet	**Scores [Independent Variable]**	**Posttest Score**	**Perception of Performance**	**Perception of Collaboration**
Construction of Shared Knowledge	N-Gram	.03	-.01	.02
	LIWC-category	.09*	.09	-.05
	Human	.11*	-.03	-.11
Negotiation/ Coordination	N-Gram	-.08	.12	.03
	LIWC-category	.03	.17	.12
	Human	-.20	.01	.11
Maintaining Team Function	N-Gram	.01	.01	.08
	LIWC-category	.02	.10	.05
	Human	-.24	.16	.32**

Note. * $p < .10$; ** $p < .05$

4.5 Effect of Transcription Errors

We investigated the effect of transcription errors (see Section 2.4) on prediction accuracy. We trained separate random forest models for the 1,114 (10%) human-transcribed utterances and their corresponding automated transcriptions. Results are show in Table 7. The human and automated transcriptions yielded similar AUROC values for construction of shared knowledge and negotiation/coordination, with the human transcriptions providing only a 2.4% and 1.5% boost in accuracy, respectively (calculated as *(human AUROC - automatic AUROC) / automatic AUROC*). This demonstrates that the errors introduced by the speech recognizer play a minimal role for these facets. For maintaining team function, the human transcriptions provide an 8.7% boost in accuracy over the automated transcriptions, suggesting transcription errors have a larger effect when facet base rates are particularly low (10%).

Table 7. AUROC for the human and automated transcriptions of 10% of the utterances.

	Human	**Automated**
Construction of Shared Knowledge	.84	.82
Negotiation/Coordination	.69	.68
Maintaining Team Function	.75	.69

5 DISCUSSION

We investigated the extent to which high-level socio-cognitive collaborative problem solving (CPS) processes (or facets) of construction of shared knowledge, negotiation/coordination, and maintaining team function, derived from a validated theoretical model of CPS [57] could be automatically modeled from spoken language during computer-supported collaborations. We used automated transcription and bag-of-n-grams approach with Random Forest models. Below we discuss our main findings along with implications, applications, limitations, and future work.

5.1 Main Findings

Our main result is that we achieved AUROC values up to .85, .77, and .77 (70%, 54% and 54% improvement over chance), for construction of shared knowledge, negotiation/coordination, and maintaining team function, respectively. This demonstrates the viability of our fully automated approach to modeling key high-level CPS processes. Indeed, we were able to achieve this performance, despite the prediction task being particularly difficult due to the open-ended nature of the collaboration. Specifically, teams were free to communicate via language, gesture, and other nonverbal channels, of which only the linguistic content was analyzed. Further, we used a fully automated pipeline, leveraging imperfect automated transcriptions. Indeed, prior to this work, it was unknown if prediction of high-level CPS facets in a videoconferencing environment was even viable. However, we demonstrated that fully-automated prediction is feasible, spoken language is a powerful modality for this task, and our modeling process is robust to imperfect data.

Importantly, our models were trained on a subset of teams and tested on other teams, thus ensuring some level of generalizability to new teams. Indeed every team is different, and the interaction dynamics can be influenced by a variety of factors such as personality [55], teammate familiarity [19], or ability [62], to name a few. Accordingly, the fact that we were able to predict the CPS facets at all is notable given these substantial individual differences. This indicates that there are high-level linguistic patterns that emerge in the context of our CPS task and that these patterns are robust to team-level (but likely not task-level)differences. Further, we included team makeup variables in our models and found no improvement in accuracy above linguistic features, again demonstrating that language patterns are diagnostic of these CPS facets.

We demonstrated that dictionary-based features (LIWC-category features) can be equally accurate as domain-specific language (n-grams). This result is particularly notable because the additional layer of abstraction (i.e., features describing language rather than the language itself) is presumably more generalizable to new domains. This is because the LIWC-category features describe categories of words related to the CPS processes and specific words are are unlikely to generalize across tasks. For example, reward words could be predictive of maintaining team function across multiple CPS tasks, whereas the use of the specific word "good" could be dependent on regional colloquial expressions of praise or framing of the task. However, this hypothesis needs empirical validation.

In a further effort to validate our model, we examined how human-coded and model-predicted facet scores predicted subjective and objective CPS outcomes after accounting for several covariates. We found that human- and LIWC-model- scores overall yielded similar predictive patterns, thus validating the LIWC computational model. However, when compared to the LIWC model, the n-gram model scores were more dissimilar to the human codes. This suggests the more generalizable LIWC features might be a better overall approach. That said, the LIWC models were not perfect as there was one notable difference in the pattern of correlations between the human and LIWC-based scores (i.e., LIWC model of maintaining team function did not predict perceptions of the collaboration while human scores did). On the other hand, the model estimates correlated with task score (though not significantly), whereas the human-codes did not, suggesting yet another discrepancy.

Overall, our results should be interpreted in light of some of our constraints. Specifically, we are modeling complex socio-cognitive processes at the utterance level, where utterances are quite short with a median of four words and 1.34 seconds duration. We also restricted our input to spoken language alone and with imperfect speech recognition. As such, the present results can be considered to be a useful lower-bound on performance, with improvement expected as additional channels are added, such as acoustic-prosodic information, facial expressions, body movements, eye gaze, and gestures.

5.2 Applications

Our work can be applied to intelligent computer interfaces that aim to monitor the ongoing collaboration, and dynamically intervene to improve CPS processes and outcomes. In particular, a collaborative interface could monitor the team's language for evidence of our facets and provide feedback

and suggestions for improvement. For example, individuals could be instructed to build on others' ideas or discuss the constraints of the task [57] if the team hits a roadblock and shared knowledge construction is determined to be low. Research is needed to identify what types of feedback and delivery mechanisms (e.g., to the individual or the team) will be most effective. Such systems could even be combined with existing feedback systems that provide feedback on low-level behaviors (Section 1.1.2), as a way of comprehensively intervening when a collaboration needs to be rerouted.

Our models can also be used offline to automatically score audio of spoken utterances without going through the time-intensive human-coding process. A key application here is the assessment and training of CPS skills, a focus of modern education [69]. Given the current accuracy scores, the models are best suited for formative assessment [53] aimed at learning rather than assessments aimed at evaluation. Model-derived estimates can be communicated as formative feedback to individuals or teams as part of an after-action review. This data can then be used to identify particular strengths and weaknesses, and to target training goals. For example, an individual with a low score of construction of shared knowledge could be encouraged to communicate their ideas with the team. Conversely, an individual with a low score of maintaining team function could be instructed to provide positive feedback to teammates more often to the extent that it is warranted.

5.3 Limitations and Future Work

Like all studies, ours has limitations. Although our models performed better than chance, they are still far from perfect. Additionally, the lack of correspondence between trained-rater- and automated-codes in predicting some of the outcome variables warrants more detailed analysis. Model performance can be improved with larger datasets and by combining language with aspects of the ongoing interaction and other multimodal signals. We expect to improve accuracy when we consider what teams are saying (language) in the context of how they are saying it (acoustic-prosodic information), what they are doing (task context), and what they feel (facial expressions and physiology). Additionally, team-level differences, such as gender makeup or personality scores, could be used as inputs to the models to improve performance.

Further, our dataset was collected at a single university, with little ethnic, age, or socioeconomic diversity. Additionally, the teams only completed a single task. Thus, our method must be verified on an extended dataset from multiple sites and tasks. We are currently collecting a multimodal dataset across multiple sites and with multiple CPS tasks in order to remedy these limitations.

5.4 Concluding Remarks

We developed fully-automated spoken language models to model three socio-cognitive processes of shared knowledge construction, negotiation/coordination, and maintaining team function during triadic, computer-supported collaborative problem solving in a manner that can generalize to new teams. Thus, we have taken a step towards the goal of intelligent collaborative interfaces that are sensitive to the unfolding collaboration process and can dynamically intervene to steer the collaboration in a more productive and satisfying direction.

6 ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF DUE 1745442) and the Institute of Educational Sciences (IES R305A170432). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

7 REFERENCES

[1] Richard Alterman and Kendall Harsch. 2017. A more reflective form of joint problem solving. *Int. J. Comput. Collab. Learn.* 12, 1 (March 2017), 9–33. DOI:https://doi.org/10.1007/s11412-017-9250-1

[2] Kathleen T Ashenfelter. 2007. *Simultaneous analysis of verbal and nonverbal data during conversation: symmetry and turn-taking*. University of Notre Dame.

[3] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv Prepr. arXiv1406.5823* 67, 1 (2014), 1–48. DOI:https://doi.org/10.18637/jss.v067.i01

[4] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the Association for Computational Linguistics 2004 on Interactive Poster and Demonstration Sessions*, 31-es. DOI:https://doi.org/10.3115/1219044.1219075

[5] Emily A. Butler and Ashley K. Randall. 2013. Emotional Coregulation in Close Relationships. *Emot. Rev.* 5, 2 (April 2013), 202–210. DOI:https://doi.org/10.1177/1754073912451630

[6] Dan Calacci, Oren Lederman, David Shrier, and Alex "Sandy" Pentland. 2016. Breakout: An Open Measurement and Intervention Tool for Distributed Peer Learning Groups. *CoRR* abs/1607.0, (2016). Retrieved from http://arxiv.org/abs/1607.01443

[7] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 873–888. DOI:https://doi.org/10.1145/2998181.2998250

[8] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* 16, 1 (March 1990), 22–29. Retrieved from http://dl.acm.org/citation.cfm?id=89086.89095

[9] Jacob Cohen. 2003. *A power primer*. American Psychological Association, Washington, DC, US.

[10] Sarah D'Angelo and Andrew Begel. 2017. Improving Communication Between Pair Programmers Using Shared Gaze Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 6245–6290. DOI:https://doi.org/10.1145/3025453.3025573

[11] Rick Dale, Riccardo Fusaroli, Nicholas D Duran, and Daniel C Richardson. 2013. The self-organization of human interaction. In *Psychology of learning and motivation*. Elsevier, 43–95.

[12] Jean Decety and Margarita Svetlova. 2012. Putting together phylogenetic and ontogenetic perspectives on empathy. *Dev. Cogn. Neurosci.* 2, 1 (2012), 1–24. DOI:https://doi.org/https://doi.org/10.1016/j.dcn.2011.05.003

[13] Dejana Diziol, Erin Walker, Nikol Rummel, and Kenneth R Koedinger. 2010. Using Intelligent Tutor Technology to Implement Adaptive Support for Student Collaboration. *Educ. Psychol. Rev.* 22, 1 (March 2010), 89–102. DOI:https://doi.org/10.1007/s10648-009-9116-9

[14] Heather A Faucett, Matthew L Lee, and Scott Carter. 2017. I Should Listen More: Real-time Sensing and Feedback of Non-Verbal Communication in Video Telehealth. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (December 2017), 44:1--44:19. DOI:https://doi.org/10.1145/3134679

[15] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 31–41.

[16] Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vis. Comput.* 27, 12 (2009), 1775–1787.

[17] Arthur C Graesser, Stephen M Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W Foltz, and Friedrich W Hesse. 2018. Advancing the Science of Collaborative Problem Solving. *Psychol. Sci. Public Interes.* 19, 2 (2018), 59–92. DOI:https://doi.org/10.1177/1529100618808244

[18] Patrick Griffin, Esther Care, and Barry McGaw. 2012. The Changing Role of Education and Schools. In *Assessment and Teaching of 21st Century Skills*, Patrick Griffin, Barry McGaw and Esther Care (eds.). Springer Netherlands, Dordrecht, 1–15. DOI:https://doi.org/10.1007/978-94-007-2324-5_1

[19] Deborah H. Gruenfeld, Elizabeth A. Mannix, Katherine Y. Williams, and Margaret A. Neale. 1996. Group composition and decision making: How member familiarity and information distribution affect process and performance. *Organ. Behav. Hum. Decis. Process.* 67, 1 (1996), 1–15. DOI:https://doi.org/10.1006/obhd.1996.0061

[20] Carl Gutwin, Scott Bateman, Gaurav Arora, and Ashley Coveney. 2017. Looking Away and Catching Up: Dealing with Brief Attentional Disconnection in Synchronous Groupware. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 2221–2235. DOI:https://doi.org/10.1145/2998181.2998226

[21] J A Hanley and B J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36. DOI:https://doi.org/10.1148/radiology.143.1.7063747

[22] Jiangang Hao, Lei Chen, Michael Flor, Lei Liu, and Alina A von Davier. 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Res. Rep. Ser.* 2017, 1 (2017), 1–9. DOI:https://doi.org/10.1002/ets2.12184

[23] Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. A Framework for Teachable Collaborative Problem Solving Skills. In *Assessment and Teaching of 21st Century Skills: Methods and Approach*, Patrick Griffin and Esther Care (eds.). Springer Netherlands, Dordrecht, 37–56. DOI:https://doi.org/10.1007/978-94-017-9395-7_2

[24] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Trans. Multimed.* 12, 6 (2010), 563–575.

[25] Melvyn J Hunt. 1990. Figures of merit for assessing connected-word recognisers. *Speech Commun.* 9, 4 (1990), 329–336. DOI:https://doi.org/https://doi.org/10.1016/0167-6393(90)90008-W

[26] Kerttu Huttunen, Heikki Keränen, Eero Väyrynen, Rauno Pääkkönen, and Tuomo Leino. 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Appl. Ergon.* 42, 2 (2011), 348–357.

[27] Hirofumi Inaguma, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2016. Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue. In *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction*, 11–15.

[28] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2018. Analyzing Gaze Behavior and Dialogue Act During Turn-taking for Estimating Empathy Skill Level. In *Proceedings of the 2018 on International Conference on Multimodal Interaction* (ICMI '18), 31–39. DOI:https://doi.org/10.1145/3242969.3242978

[29] Patrick Jermann and Marc-Antoine Nüssli. 2012. Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1125–1134.

[30] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst.* 3, 2 (August 2013), 12:1--12:30. DOI:https://doi.org/10.1145/2499474.2499481

[31] Janice R Kelly and Sigal G Barsade. 2001. Mood and emotions in small groups and work teams. *Organ. Behav. Hum. Decis. Process.* 86, 1 (2001), 99–130.

[32] Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1 (2017), 559–563.

[33] Rivka Levitan, Agustin Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL HLT '12), 11–19.

[34] Dekang Lin. 1998. Extracting collocations from text corpora. In *First workshop on computational terminology*, 57–63.

[35] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, 17–24.

[36] Moira McGregor and John C Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 2208–2220. DOI:https://doi.org/10.1145/2998181.2998335

[37] Christian Müller, Barbara Großmann-Hutter, Anthony Jameson, Ralf Rummer, and Frank Wittig. 2001. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In *International Conference on User Modeling*, 24–33.

[38] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces*, 153–164.

[39] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (ICMI '18), 14–20. DOI:https://doi.org/10.1145/3242969.3243027

[40] Fumio Nihei, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting Influential Statements in Group Discussions Using Speech and Head Motion Information. In *Proceedings of the 16th International Conference on Multimodal Interaction* (ICMI '14), 136–143. DOI:https://doi.org/10.1145/2663204.2663248

[41] José C. Pinheiro and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. Springer Verlag, New York, NY.

[42] Martin F. Porter. 2001. Snowball: A Language for Stemming Algorithms.

[43] David Nadler Prata, Ryan S J d Baker, Evandro d B Costa, Carolyn P Rosé, Yue Cui, and Adriana M J B De Carvalho. 2009. Detecting and Understanding the Impact of Cognitive and Interpersonal Conflict in Computer Supported Collaborative Learning Environments. *Int. Work. Gr. Educ. Data Min.* (2009).

[44] Jeremy Roschelle and Stephanie D Teasley. 1995. The Construction of Shared Knowledge in Collaborative Problem Solving. In *Computer Supported Collaborative Learning*, 69–97.

[45] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Int. J. Comput. Collab. Learn.* 3, 3 (September 2008), 237–271. DOI:https://doi.org/10.1007/s11412-007-9034-0

[46] Yigal Rosen. 2015. Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *Int. J. Artif. Intell. Educ.* 25, 3 (2015), 380–406. DOI:https://doi.org/10.1007/s40593-015-0042-3

[47] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. 2018. CoCo: Collaboration Coach for Understanding Team Dynamics During Video Conferencing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (January 2018), 160:1--160:24. DOI:https://doi.org/10.1145/3161186

[48] C Schlösser, A Harrer, and A Kienle. 2018. Supporting Dyadic Chat Communication with Eye Tracking Based Reading Awareness. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 149–151. DOI:https://doi.org/10.1109/ICALT.2018.00042

[49] Christian Schlösser. 2018. Towards concise gaze sharing. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 78.

[50] Bertrand Schneider and Roy Pea. 2013. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *Int. J. Comput. Collab. Learn.* 8, 4 (2013), 375–397.

[51] Julian Schulze and Stefan Krumm. 2017. The "virtual team player": A review and initial model of knowledge, skills, abilities, and other characteristics for virtual collaboration. *Organ. Psychol. Rev.* 7, 1 (February 2017), 66–95. DOI:https://doi.org/10.1177/2041386616675522

[52] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and others. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One* 8, 9 (2013), e73791.

[53] Valerie J. Shute. 2008. Focus on Formative Feedback. *Rev. Educ. Res.* 78, 1 (March 2008), 153–189. DOI:https://doi.org/10.3102/0034654307313795

[54] Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, 13–20.

[55] Angela E. B. Stewart and Sidney K D'Mello. 2018. Connecting the Dots Towards Collaborative AIED: Linking Group Makeup to Process to Learning. In *International Conference on Artificial Intelligence in Education*, 545–556.

[56] Kristin Stoeffler, Yigal Rosen, Maria Bolsinova, and Alina von Davier. 2018. Gamified Assessment of Collaborative Skills with Chatbots. . 343–347. DOI:https://doi.org/10.1007/978-3-319-93846-2_64

[57] Chen Sun, Valerie Shute, Angela E.B. Stewart, Jade Yonehiro, Nicholas Duran, and Sidney K. D'Mello. Toward a generalized competency model of collaborative problem solving. In Press *Computers and Education.*

[58] Yla R Tausczik and James W Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* 29, 1 (2010), 24–54. DOI:https://doi.org/10.1177/0261927X09351676

[59] F. B. M. de Waal. 2012. The Antiquity of Empathy. *Science (80-.).* 336, 6083 (May 2012), 874–876. DOI:https://doi.org/10.1126/science.1220999

[60] Mary Webb and David Gibson. 2015. Technology enhanced assessment in complex collaborative settings. *Educ. Inf. Technol.* 20, 4 (December 2015), 675–695. DOI:https://doi.org/10.1007/s10639-015-9413-5

[61] Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2015. Virtual Teams in Massive Open Online Courses. In *Artificial Intelligence in Education*, 820–824.

[62] Ian A.G. G Wilkinson and Irene Y.Y. Y Fung. 2002. Small-group composition and peer effects. *Int. J. Educ. Res.* 37, 5 (2002), 425–447. DOI:https://doi.org/10.1016/S0883-0355(03)00014-4

[63] Nancy Yao, Jeff Brewer, Sarah D'Angelo, Mike Horn, and Darren Gergle. 2018. Visualizing Gaze Information from Multiple Students to Support Remote Instruction. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, LBW051.

[64] Bo Yin and Fang Chen. 2007. Towards automatic cognitive load measurement from speech analysis. In *International Conference on Human-Computer Interaction*, 1011–1020.

[65] Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve W J Kozlowski, and Hayley Hung. 2018. The I in Team: Mining Personal Social Interaction Routine with Topic Models from Long-Term Team Data. In *23rd International Conference on Intelligent User Interfaces*, 421–426.

[66] Code Studio. Retrieved April 1, 2018 from https://studio.code.org/s/mc/stage/1/puzzle/1

[67] IBM. Retrieved May 2, 2018 from https://www.ibm.com/watson/services/speech-to-text/

[68] Scikit Learn. Retrieved May 3, 2018 from https://github.com/scikit-learn/scikit-learn

[69] 2015. *PISA 2015 Collaborative Problem Solving Framework.*

Received April 2019; revised June 2019; accepted August 2019.