Disfluent but effective? A quantitative study of disfluencies and conversational moves in team discourse

Felix Gervits[1] **Kathleen Eberhard**[2] **Matthias Scheutz**[1]
[1]Tufts University, Human-Robot Interaction Laboratory, Medford, MA
[2]University of Notre Dame, Department of Psychology, Notre Dame, IN
{felix.gervits, matthias.scheutz}@tufts.edu
eberhard.1@nd.edu

Abstract

Situated dialogue systems that interact with humans as part of a team (e.g., robot teammates) need to be able to use information from communication channels to gauge the coordination level and effectiveness of the team. Currently, the feasibility of this end goal is limited by several gaps in both the empirical and computational literature. The purpose of this paper is to address those gaps in the following ways: (1) investigate which properties of task-oriented discourse correspond with effective performance in human teams, and (2) discuss how and to what extent these properties can be utilized in spoken dialogue systems. To this end, we analyzed natural language data from a unique corpus of spontaneous, task-oriented dialogue (CReST corpus), which was annotated for disfluencies and conversational moves. We found that effective teams made more self-repair disfluencies and used specific communication strategies to facilitate grounding and coordination. Our results indicate that truly robust and natural dialogue systems will need to interpret highly disfluent utterances and also utilize specific collaborative mechanisms to facilitate grounding. These data shed light on effective communication in performance scenarios and directly inform the development of robust dialogue systems for situated artificial agents.

1 Introduction

Effective coordination is fundamental to teamwork in many fields, particularly for teams working under stress. Though a variety of team structures exist, actions teams are among the most demanding, due to the need for the teammates to engage in goal-oriented, interdependent tasks, and to dynamically adapt their decision-making, communication, and planning strategies (Serfaty et al., 1993; Sundstrom et al., 2000). Recently, action teams have begun to incorporate artificial agents in mixed-initiative roles (see Sycara and Sukthankar (2006) for a thorough review). Such teams are largely employed in performance scenarios (e.g., search and rescue missions, military squads, surgical teams), and require strong team communication to achieve complex objectives in a time-sensitive context. Among the many elements needed for coordination in human-agent teams, perhaps the most important for team success is establishing *common ground*. Common ground is a mutual understanding between teammates, involving shared knowledge of the task environment, goals, perspectives, and other factors. For common ground to be established, it is important that teammates not only share information, but also ensure that the information was understood the way it was intended. This process, known as *grounding*, results in a mutual recognition by both parties of the shared information being a part of their common ground and forms the basis for coordination in both dialogue and action (Clark, 1996).

1.1 Task-oriented remote dialogue

Though there is no objective way to measure common ground, evidence for it may be found in the team communication channels. However, grounding in task-oriented remote communication is complicated by a number of factors, including time pressure/workload (Entin and Serfaty, 1999; Khawaja et al., 2012; Urban et al., 1996), mode of interaction (Clark and Brennan, 1991; Doherty-Sneddon et al., 1997; Krauss and Weinheimer, 1966), and team structure (Bortfeld et al., 2001; Clark and Krych, 2004). Given the difficulty of grounding exchanges in task-oriented dialogue, it is not surprising that communication channels in remotely-communicating action teams are very noisy, often containing features such as disfluency, overlapping speech, and ambiguity.

Disfluencies are particularly interesting because they are very common in spontaneous speech, and have been implicated in various interpersonal and cognitive functions (Nicholson et al., 2003). One view holds that disfluencies are noise resulting from increased production difficulty due to cognitive workload. Several studies have found support for this view by showing that disfluency rates tend to increase with higher workload (Berthold and Jameson, 1999; Lindström et al., 2010). Another position is that disfluencies may not be solely due to workload, but may reflect underlying coordination processes such as monitoring one's addressee (Clark and Krych, 2004) or soliciting help in the dialogue (Bortfeld et al., 2001). In support of this view are studies showing that speakers detect and utilize disfluencies to help process surrounding speech (Brennan and Schober, 2001), resolve reference ambiguities (Arnold et al., 2007), hold the conversational floor (Smith and Clark, 1993), improve recall (Corley et al., 2007), and mark discourse structure (Swerts, 1998). Barr (2001) has likened disfluencies to a form of "vocal gestures" due to their ability to provide insight into a speaker's metacognitive state. Despite these findings, no prior studies have examined the role of speech disfluencies with regard to performance in an unscripted collaborative task. Thus, it remains unclear whether the interpersonal benefit of disfluency overrides the cognitive drawback also associated with disfluent speech. Moreover, most of the existing literature on the benefit of disfluency deals with filled pauses, whereas relatively little is known about other types, such as self-repairs. It is also unclear what types of coordination strategies (if any) successful teams use to overcome the various grounding constraints associated with task-oriented communication, and how these strategies interact with a team's ability to establish common ground. Our study was designed to address these gaps in the literature.

1.2 Present study

The present study consists of a quantitative investigation of factors that influence effective grounding and performance in remotely-communicating action teams. Given that grounding is often constrained by factors such as coordination strategy, speaker role, and time pressure/workload, our aim was to explore how these factors interact with team communication and, ultimately, performance. The data consisted of the linguistic- and dialogue-level annotations in the Cooperative Remote Search Task (CReST) corpus (Eberhard et al., 2010) obtained from unscripted communication between a (remotely-located) director and a searcher, who was located in an indoor environment and collaboratively performed a number of tasks involving objects in the environment (see Section 2 for more information). Since the task was specifically designed to simulate the structure of action teams in which a robot may play a vital role (e.g., urban search and rescue), the results can inform our understanding of the kinds of communication and coordination strategies that artificial agents will need to adopt to be effective partners in these hierarchical, mixed-initiative teams.

1.3 Predictions

One specific prediction is that effective teams would be those in which the director plays a greater role in managing the task, and searchers are the more receptive party. Additionally, effective teams will minimize *joint collaborative effort* (Clark and Wilkes-Gibbs, 1986) by establishing common ground with respect to objects and locations in the environment. Evidence for this would be the use of various communication strategies, including: establishing shared referents and shorthand conventions (Clark and Wilkes-Gibbs, 1986), completing one another's utterances (Clark and Schaefer, 1989), taking a partner's perspective (Brennan et al., 2010), and breaking up longer utterances into installments (Clark and Brennan, 1991).

Evidence of collaborative dialogue may also manifest in increased disfluency rate. Though some disfluent speech may be expected due to the difficult nature of the task, an increase in self-repairs may also indicate that a speaker is self-monitoring and adjusting their speech for clarity and accuracy (Clark and

Krych, 2004). In this way, self-repairs may serve to minimize collaborative effort by fixing a problematic utterance before one's partner needs to intervene (Levelt, 1983). This can minimize the number of dialogue turns and lead to more efficient exchanges. Thus, we predict that effective teams would make more disfluencies due to the need to plan and coordinate at higher speeds. Disfluencies would be mainly self-repairs caused by monitoring one's listener and dynamically adjusting one's speech to aid comprehension.

2 Method

The CReST corpus (Eberhard et al., 2010) was used to evaluate speech patterns in 10 teams (20 individuals) of people performing a collaborative, remote, search task. Approximately 8 minutes of language data were extracted for each team. Contained in the corpus is dialogue that occurred before and after a time-limit warning which allows us to examine the effects of time pressure. The corpus also provides an objective measure of the pairs' task performance, which we used to operationalize "effectiveness of communication". Additionally, the members of each team had asymmetrical roles, which we included as an additional factor. Lastly, the corpus was annotated for various linguistic and dialogue events in the speech, including conversational moves and disfluencies (see Eberhard et al. (2010) for additional details about the corpus).

2.1 Task description

The members of each pair were randomly assigned to the *Director* role and *Searcher* role. The director was seated in front of a computer that displayed a floor plan map of the search environment and wore a headset for remotely communicating with the searcher. The searcher also wore a headset and was situated in the search environment which consisted of a hallway and 6 connected office rooms. Neither was familiar with the environment. Distributed throughout the environment were 8 blue boxes, each with three colored blocks, 8 empty green boxes (numbered 1-8), 8 empty pink boxes, and a cardboard box that was at the furthest point from entrance at the end of a hallway. Some of the colored boxes were partially hidden behind a door, on a chair, under a table, etc.

The pairs were informed that the director's map showed the locations of all the boxes except the green ones and that the locations of some of the blue boxes were inaccurate. They were told that the searcher was to retrieve the cardboard box, put the blue blocks from the blue boxes into it, and report the locations of the green boxes to the director, who was to mark them on the map by dragging green icons numbered 1-8. They were told that instructions for the pink boxes would be given to them later. Five minutes into the task, the director's communication with the searcher was put on hold and the director was told that each blue box contained a yellow block which was to be put into each of the pink boxes. To examine effects of time pressure, the director also was told that they had 3 minutes to complete all of the objectives, and a timer that counted down the 3 minutes was displayed next to the map.

2.2 Disfluency annotation

Disfluencies were coded according to the HCRC Disfluency Coding Manual (Lickley, 1998), which includes categories for prolongations, pauses (filled and silent), and self-repairs: *repetitions* (e.g., "L-look in the box"), *substitutions* (e.g., "the pink- uh, blue box"), *insertions* (e.g., "go into the room- the nearby room") and *deletions* (e.g., "we don't have- let's hurry up"). Disfluency rates were calculated for each participant as a proportion per every 100 words. *Speech rate* (words per minute, or w.p.m.) and *mean length of utterance* (average number of words per turn at talk, or MLU) also were calculated.

All annotations were carried out using the open-source EXMARaLDA Partitur-Editor (Schmidt and Wörner, 2009). For extracting disfluency data from the annotated files, we used a custom-built search tool called DeepSearch9[1].

[1] Our custom search tool, DeepSearch9, will be made available for research purposes

2.3 Dialogue annotation

The transcribed utterances were hand-annotated for type of conversational move using Carletta et al. (1997)'s scheme. *Initiation* moves include Instruct, Explain, Wh- and Yes/No questions. Two other Initiation moves are subcategories of Yes/No questions, namely, Check and Align. Checks seek confirmation that one has correctly understood what the partner recently said, often by repeating or paraphrasing the partner's utterance. Aligns explicitly request confirmation that a partner has understood what was just said and is ready to move on. They typically are in the form of an "okay?" or "right?" appended to the end of an Instruct or Explain move. *Response* moves include Acknowledge, Wh-, Yes- and No- Replies. Utterance-initial "okays" and "alrights" were coded as *Ready* moves; they serve as a preparation for the following initiation move (e.g., "*Okay*, now go into the next room"). The rates of producing each type of move were calculated by dividing them by the total number of utterances.

2.4 Team effectiveness

Performance was scored with respect to the number of colored boxes whose task was completed, with a maximum score of 24. The average score was 9.9 (range 1 - 19) and the median was 8. The median score was used to divide the 10 teams into an *effective* and *ineffective* group with average scores of 14.8 (S.D. = 4.0) and 5.0 (S.D. = 2.5), respectively.

3 Results

To test our hypotheses of the factors that influence effective team communication, we examined differences in disfluency rate and dialogue moves between teams in the effective and ineffective performance groups. Time pressure and speaker role were used as factors in the analysis. Some relevant dialogues from the corpus are also discussed below to show the various communication and grounding strategies that teams used.

3.1 Grounding strategies

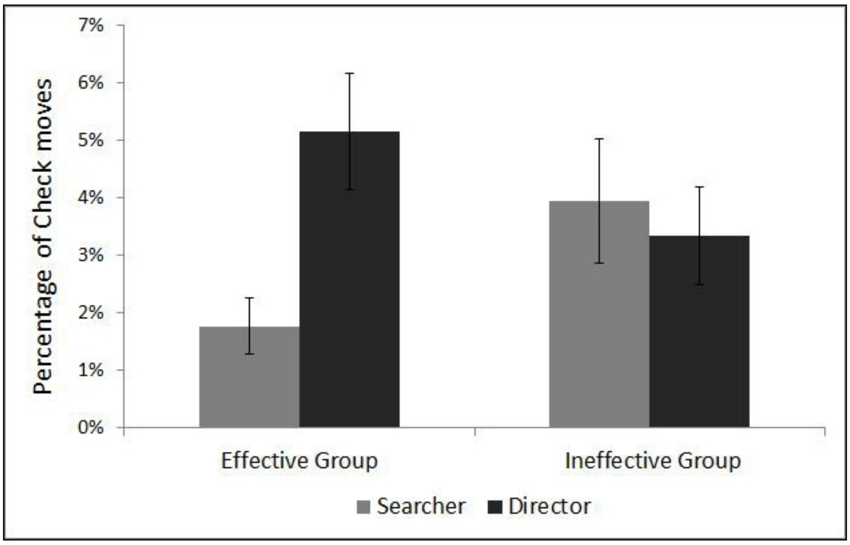

Figure 1: Group x Speaker interaction for Check moves. Error bars represent standard error of the mean.

First, we analyzed group differences in the dialogue moves to test our hypothesis that team effectiveness was related to successful grounding and collaboration. The rates of types of dialogue moves were analyzed with 2x2x2 mixed ANOVAs with Time Pressure and Speaker as within-subjects factors and Group as a between-subjects factors. There was a Group x Speaker interaction for *Check* moves ($F(1,32) = 7.053$, $p = .012$), with effective directors producing more than the ineffective directors, and ineffective searchers producing more than the effective searchers (see Fig. 1). The 3-way interaction was also significant in the analysis of the *Ready* moves ($F(1,32) = 4.657$, $p = .039$). Effective directors' rate of *Ready* moves was higher than the effective searchers' with and without time pressure, whereas ineffective directors' rate of *Ready* moves decreased to the level of the ineffective searchers' under time pressure (see Fig. 2). Together, the results support our prediction that, compared to ineffective directors, effective directors sought confirmation of their searcher's understanding more often (*Check* moves) and maintained consistent control over the dialogue structure (*Ready* moves).

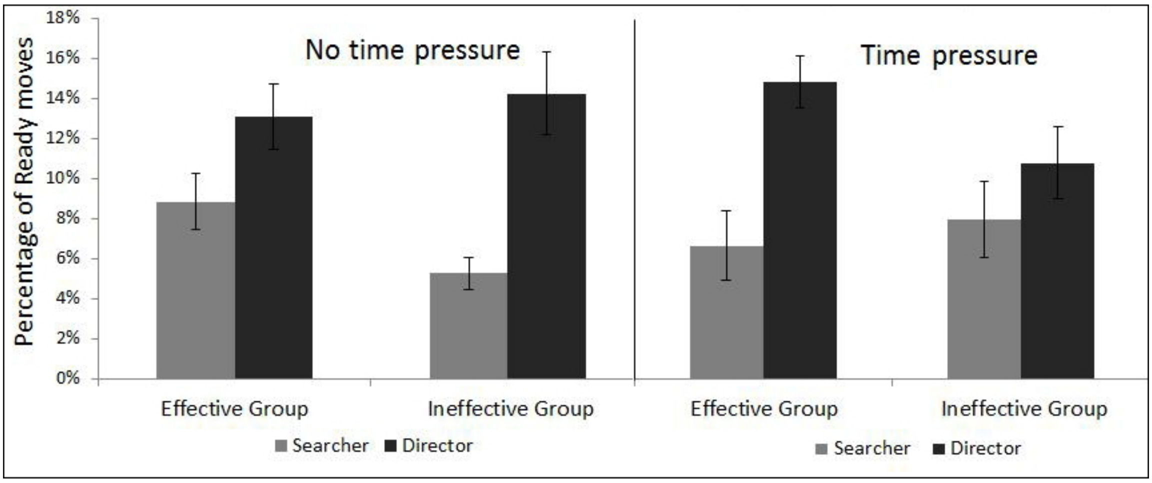

Figure 2: Group x Speaker x Time Pressure interaction for Ready moves. Error bars represent standard error of the mean.

One particular communication strategy used by effective teams was grounding of referents - particularly in room labeling. Since the rooms were not labeled on the map or in the environment, some of the teams developed and utilized a shorthand way to describe them to each other. Consider the following example from an effective team:[2]

D: Okay you're going into the next room
S: Room two
D: Room two, we'll say room two
S: Alright

This is an example of the searcher using a *replacement* (Clark and Wilkes-Gibbs, 1986) to ground a referent. The director initially said "next room", but the searcher proposed "room two" instead. Notice how the director agreed to this shorthand label, and then this was again confirmed by the searcher. This contribution served to establish "room two" as part of the team's common ground. They were able to use this later to minimize joint effort, as in:

S: I'm walking back into [pause] room two
D: Okay
—
S: I'm going back into room one
D: Okay room one, like the very first starting room?
S: Yeah
D: Okay

Here, the label "room one" was used by the searcher, although this was not previously established. The director initiated a *Check* utterance to confirm that they were talking about the same room. After

[2] In the dialogue examples, hyphens (-) indicate repaired segments, colons (:) indicate prolongations, and commas (,) indicate brief silent pauses, with longer pauses contained in brackets. For readability, the director will be referred to as male and the searcher as female in this and all subsequent dialogue examples.

confirmation by the searcher, the grounding was successful and the interaction continued along. Ineffective teams exhibited difficulty grounding referents as seen in the following example:

S: So green number 3 [pause] that was um
D: In the booth
S: Yeah that was in the- [pause] the little room
D: Where in the booth? Where?
S: It wasn't in the big room, it was in the little room, it was right next to the blue box on the chair

Here, the director referred to "the booth", while the searcher referred to a "little room". Since they did not establish common terminology for these referents, the team lost efficiency by taking additional turns.

Additional conversational turns were not always signs of inefficient coordination, however. For example, effective teams facilitated grounding by presenting a task subgoal in installments (Clark and Brennan, 1991) as illustrated here:

D: If you: turn around go out of that room
S: Okay
D: Straight in front of you should be a chair
S: Yes
D: At a table, there's a blue box there
S: Yes
D: Okay, get that

The searcher's acknowledgement following each installment allowed the director to continue on. This strategy also identifies the particular point of grounding difficulty, which is not possible when an entire subgoal is communicated in a single complex turn as illustrated by the following example from an ineffective team:

D: If you look completely straight- straight- straight [pause] like keep walking straight before you even hit the wall, there should be some shelving it looks like. Open the blue box there
S: Wait w- where- where? Sorry {laughs}

Overall, these results suggest that effective teams were better at coordinating their actions by using various dialogue moves and communication strategies to enhance grounding. Effective directors also played a central role in managing the task responsibilities, and sustained this initiative even under time pressure.

3.2 Disfluencies as collaborative tools

To test our hypothesis that disfluencies serve collaborative functions, we examined group differences in disfluency rate. A MANOVA was conducted on the rates of the four types of self-repairs, with Group and Time Pressure as factors. There was a significant effect of Group (F(4,33) = 2.787, p = .042) on rates of self-repair disfluencies (see Fig. 3): *Insertions* (F(1,36) = 4.292, p = .046), *Deletions* (F(1,36) = 4.414, p = .043), and a trending effect for *Substitutions* (F(1,36) = 2.826, p = .101). In all cases, the effective group had higher disfluency rates, which was not due to longer utterances because the groups' MLU did not differ ($F < 1$). For hesitation disfluencies, a MANOVA conducted on prolongations, filled- and silent-pauses found no effect of performance group on these disfluency measures ($F < 1$). This was expected because these disfluencies (especially filled pauses) were the most common in the corpus, and did not exhibit speaker differences in previous analyses (Nicholson et al., 2003). Overall, the finding that self-repairs increased for the effective teams supports our hypothesis that these types of disfluencies are not solely due to production difficulty, but rather may serve an interpersonal function.

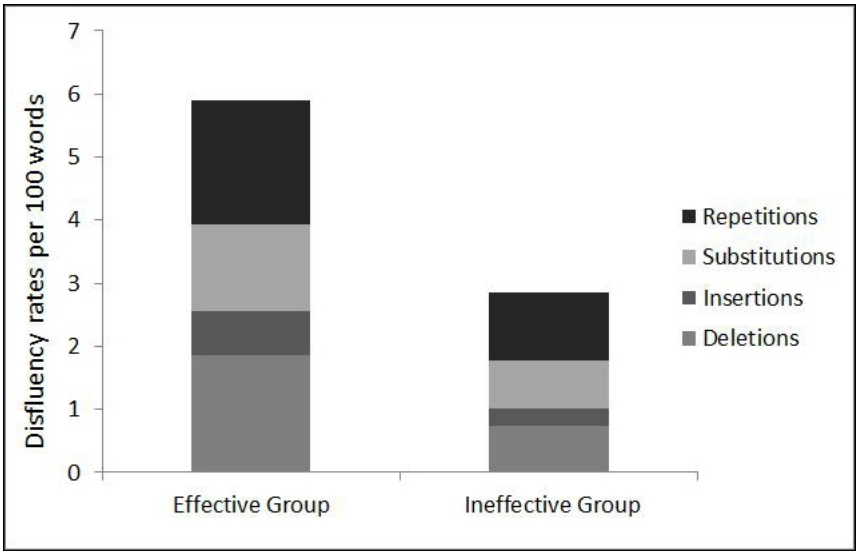

Figure 3: Group effect for disfluency rate

Examples from the corpus further show how disfluencies were used by effective teams as collaborative tools to enhance grounding. Consider the following dialogue exchange from one of the top performing teams under time pressure:

S: Yeah, there's one all the way to the right
D: Okay, you should see- you should [pause] be- go to that end hallway

The director's mixed substitution/deletion self-repair served to make the utterance more accurate and clear for the searcher. Since the director was unsure of exactly what the searcher was seeing, he repaired his utterance and changed it to an instruction that did not provide false information, and that was simpler to understand. This is in line with prior evidence (e.g., Clark and Wilkes-Gibbs (1986)) showing that self-repairs increase when people accommodate their partner's perspective.

Due to the fast-paced nature of the task, some searchers tended to "think out loud" and update their directors with new information as they obtained it. This naturally led to an increase in disfluency, as in the following example from an effective team:

S: But I'm out of- I'm out of- uh [pause] actually, there's a blue one to my left
D: Yes there should be- yes pick that up

In this exchange, the searcher started to say that she is out of blocks, but then changed this mid-sentence (via deletion disfluency) to share new information ("there's a blue one to my left"). Similarly, the director used a substitution to confirm the prior utterance ("yes there should be") and also as a new instruction ("yes pick that up"). The disfluencies here maximized the information that each speaker shared in a single turn, leading to increased efficiency under time pressure.

Effective teams were also able to resolve reference ambiguities through the use of disfluencies, as in the following example:

D: There's also one in the second- [pause] uh, we only have three minutes to do this, okay
S: Okay, second cubicle I got that

Here, the searcher was able to predict that the director was referring to the cubicle, even though this was part of the deleted reparandum, and the word "cubicle" was not even explicitly uttered. The silent pause combined with the non-lexical filler "uh" may have signaled that the director was referring to an object, namely, the cubicle. Despite the fact that this segment of the utterance was deleted, the searcher was still able to resolve the intended referent by drawing on the mutual knowledge that both teammates share a similar floor plan of the room.

For effective teams in the study, the benefit of clarifying an utterance may have outweighed the cost of disfluent speech because it is cheaper for a person to repair their own speech rather than to have a partner do it for them (Levelt, 1983). In cases where an error was not self-repaired (such as the miscommunication of an important goal), this could have led to misinformation resulting in confusion or loss of efficiency. As a whole, these data suggest that self-repairs served as collaborative tools in the discourse, and were utilized by effective teams to enhance coordination and performance.

4 Discussion

4.1 Summary of results

Overall, the results of our investigation revealed a number of novel effects of disfluencies and dialogue moves in task-oriented remote communication. In our analysis of dialogue moves, we found that the best performing directors used more Check utterances than the ineffective directors (see Fig. 1). Check utterances are used as a means to gauge a partner's understanding, and are especially useful for reducing uncertainty when interlocuters are remotely separated and need to establish shared knowledge. Effective directors also produced more Ready utterances under time pressure than ineffective directors (see Fig. 2), indicating that they showed receptiveness to their partner and maintained control over guiding the team through the changing task conditions. We also observed that more effective teams produced twice as many self-repair disfluencues as the less effective teams. Although previous studies have demonstrated some effects of disfluency on dialogue (Arnold et al., 2007; Brennan and Schober, 2001), ours is the first to link these effects with improved performance in a joint task involving spontaneous speech. Our results indicate that effective teammates were monitoring their own speech and repairing it for clarity and accuracy in order to accommodate their partner's perspective and facilitate grounding. In this way, disfluencies suggest a greater team awareness, which may have contributed to improved performance (see Fig. 3). Overall, the ability to establish common ground through collaborative exchanges seemed to be the key factor in effective team performance. Effective teams were able to accomplish this through particular communication strategies that involved: self-monitoring one's speech for clarity (e.g., self-repairs), being responsive to one's teammate (e.g., Ready moves), as well as monitoring them for understanding (e.g., Check moves).

4.2 Application to spoken dialogue systems

Our results shed light on task-oriented communication and directly inform the development of robust dialogue systems that rely on interactive, collaborative mechanisms. It is important that such systems are able to utilize the information contained in team discourse to gain insight into speakers' cognitive and performance states, and to make better predictions about the course of the interaction. Specifically, the finding that self-repairs are strong indicators of collaborative processes and are increasingly utilized in effective teams, suggests that the detection and interpretation of speech disfluency can be of great benefit to dialogue systems. While there have been numerous approaches to automated disfluency detection (many of which used statistical models trained on the Switchboard corpus, e.g., Qian and Liu (2013)), our findings suggest that these approaches lack several specific components needed for use in robust dialogue systems, including: (1) incremental, on-line processing, (2) identifying the function of the repair, (3) use of the repair to make predictions about subsequent dialogue, (4) use of additional non-linguistic information in the model (e.g., speaker role, cognitive state, etc.), and (5) generalizing to domains other than two-party telephone conversations. While some existing systems attempt to solve these problems individually (e.g., domain generality: Georgila et al. (2010); repair function: Hough et al. (2013); incrementality: Zwarts et al. (2010)), no current approach combines these elements into a unified system. Such an integrated approach is important going forward because it will allow systems to handle the kinds of natural utterances that people make in task-oriented discourse.

Importantly, our findings also highlight the need for dialogue systems to handle incremental input. Given the abundance of time-sensitive information that can be extracted from discourse channels (disfluency, turn-taking, back-channel feedback, etc.), this is an important next step to improve dialogue processing. Though some progress has recently been made on dialogue managers that can handle incremental semantic input (Buß et al., 2010; Schlangen et al., 2009), there has been little to no work on integrating these mechanisms with the planning and reasoning (not to mention vision and motor) capabilities of artificial agents that coordinate their actions with humans. This is a necessary step in order to test the benefit of the collaborative mechanisms our study revealed, and to advance the state of the art of spoken dialogue systems.

4.3 Future directions

Future work will be necessary on both the experimental and computational ends. Future studies will need to evaluate larger samples that are carefully controlled for variables that might influence performance (e.g., gender, age, familiarity, etc.). In addition, workload will need to be objectively measured in order to determine if the effects of time pressure were experienced differently by teams of varying performance. An objective workload measure will also allow for investigation into the *cause* of speech disfluency, specifically in terms of whether it is due to task demands or to intentional signaling (Nicholson et al., 2003).

On the computational end, our results support the development of spoken dialogue systems that can track various dimensions of "team cohesion" based on discourse measures. Such systems can be used in team training exercises in order to identify when a team is under-performing, to find areas of weakness, and to improve overall coordination. An artificial agent with such capacity can also take appropriate actions to repair a problematic interaction through a range of strategies (e.g., *Check* and *Ready* moves, frequent acknowledgments, establishing shared referents, etc.) that we and others found to minimize joint collaborative effort. As a whole, these abilities would improve on the capabilities of existing dialogue systems, and would enable artificial agents to function as more effective teammates.

5 Conclusions

This first study connecting spontaneous speech, dialogue, and task performance supports previous literature and introduces novel findings about the effects of self-repair disfluencies and dialogue moves on coordination and performance in remotely-communicating action teams. The best-performing teams in our study utilized a variety of communication strategies to enhance coordination, including monitoring for understanding, perspective-taking, self-repairs and others. Importantly, we showed that these strategies can be identified by discourse properties available in team communication channels. Applying these empirical results to spoken dialogue systems is an important future direction that can lead to more natural and improved human-agent interaction. By better understanding effective team communication we can design more robust systems that take advantage of the kinds of interactive, collaborative mechanisms inherent in dialogue, to improve coordination and performance in all kinds of human teams.

6 Acknowledgments

This work was supported in part by grants N00014-07-1-1049, N00014-11-1-0493, and N00014-14-1-0149 from the US Office of Naval Research. We would also like to thank the anonymous reviewers for their helpful feedback and suggestions.

References

Jennifer E. Arnold, Carla L. Hudson Kam, and Michael K. Tanenhaus. 2007. If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):914–930.

Dale Barr. 2001. Trouble in mind: paralinguistic indices of effort and uncertainty in communication. In C. Cavé, I. Guaïtella, and S. Santi, editors, *Oralité et gestualité: interactions et comportements multimodaux dans la communication: actes du colloque ORAGE 2001*, pages 597–600. Harmattan, Paris.

Andre Berthold and Anthony Jameson. 1999. Interpreting symptoms of cognitive load in speech input. In Judy Kay, editor, *UM99, User Modeling: Proceedings of the Seventh International Conference*, pages 235–244. Springer Wien New York.

Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

Susan E. Brennan and Michael F. Schober. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44(2):274–296.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, 53:301–344.

Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 233–236.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C. Kowtko, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, 13:127–149.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Herbert H. Clark. 1996. *Using language*. Cambridge university press.

Martin Corley, Lucy J. MacGregor, and David I. Donaldson. 2007. Its the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'Malley, Steve Langton, Simon Garrod, and Vicki Bruce. 1997. Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(2):105–125.

Kathleen M. Eberhard, Hannele Nicholson, Sandra Kübler, Susan Gundersen, and Matthias Scheutz. 2010. The Indiana "Cooperative Remote Search Task" (CReST) corpus. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23*.

Elliot E. Entin and Daniel Serfaty. 1999. Adaptive Team Coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41(2):312–325.

Kallirroi Georgila, Ning Wang, and Jonathan Gratch. 2010. Cross-domain speech disfluency detection. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–240.

Julian Hough, Matthew Purver, et al. 2013. Modelling expectation in the self-repair processing of annotat-, um, listeners. In *Proceedings of the 17th SemDial workshop on the Semantics and Pragmatics of Dialogue*.

M Asif Khawaja, Fang Chen, and Nadine Marcus. 2012. Analysis of collaborative communication for linguistic cues of cognitive load. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(4):518–529.

Robert M. Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of personality and social psychology*, 4(3):343.

Willem Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Robin Lickley. 1998. *HCRC Disfluency Coding Manual*. Human Communication Research Centre, University of Edinburgh.

Anders Lindström, Jessica Villing, Staffan Larsson, Alexander Seward, and Cecilia Holtelius. 2010. The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In *Proceedings of the International Conference on Spoken Language Processing, Interspeech, 17-23*.

Hannele Nicholson, Ellen Gurman Bard, Rohin Lickley, Anne H. Anderson, Jim Mullin, David Kenicer, and Lucy Smallwood. 2003. The intentionality of disfluency: Findings from feedback and timing. In *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*.

Xian Qian and Yang Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of the NAACL-HLT*.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 30–37.

Thomas Schmidt and Kai Wörner. 2009. Exmaralda-creating, analyzing and sharing spoken language corpora for pragmatics research. *Pragmatics*, 19(4):565–582.

Daniel Serfaty, Elliot E. Entin, and Catherine Volpe. 1993. Adaptation to stress in team decision-making and coordination. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 37, pages 1228–1232.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38.

Eric Sundstrom, Michael McIntyre, Terry Halfhill, and Heather Richards. 2000. Work groups: From the Hawthorne studies to work teams of the 1990s and beyond. *Group Dynamics: Theory, Research, and Practice*, 4(1):44–67.

Marc Swerts. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30:485–496.

Katia Sycara and Gita Sukthankar. 2006. Literature review of teamwork models. Technical Report CMU-RI-TR-06-50, Robotics Institute, Pittsburgh, PA.

Julie M. Urban, Jeanne L. Weaver, Clint A. Bowers, and Lori Rhodenizer. 1996. Effects of workload and structure on team processes and performance: Implications for complex team decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(2):300–310.

Simon Zwarts, Mark Johnson, and Robert Dale. 2010. Detecting speech repairs incrementally using a noisy channel approach. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1371–1378.