Thesis Proposal:
Agile Facilitation for Collaborative Learning

David Adamson

February 2014

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carolyn P. Rosé, Chair
Vincent Aleven
Alexander Rudnicky
Diane Litman

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.
**Keywords:** Computer Supported Collaborative Learning, Conversational Agents, Intelligent Tutoring Systems, Automated Discourse Analysis, Accountable Talk
For all the teachers and students at Digital Harbor High School
Abstract

This proposed work investigates the use of adaptable conversational agents to scaffold online collaborative learning discussions through an approach called Academically Productive Talk (APT), or Accountable Talk. In contrast to past work on dynamic support for collaborative learning, where agents were used to elevate conceptual depth by leading students through directed lines of reasoning, this APT-based approach uses generic prompts that encourage students to articulate and elaborate their own lines of reasoning, and to challenge and extend the reasoning of their teammates. Our body of completed work integrates findings from a series of studies across content domains (biology, chemistry, engineering design), grade levels (high school, undergraduate), and facilitation strategies. The pattern of results demonstrates that APT based support for collaborative learning can significantly increase learning, but that the effect of specific APT facilitation strategies is context-specific. It appears the effectiveness of each strategy depends upon factors such as the difficulty of the material and the skill level of the learner.

As a primary contribution of this work, we plan to develop and operationalize a framework for automated, context-responsive facilitation. Through a series of in-vivo experiments and transcript analyses, we seek to verify and clarify which contextual factors are most critical when selecting suitable facilitation strategies. Prior work has shown that models learned from sequences of discourse acts can predict learning and social outcomes, even from just the early stages of an ongoing discussion, and that representations of a conversation’s evolving state hold benefit for both human and automated support. How do differences among macro-level collaborative contexts and turn-level interventions affect a group’s conversational trajectory? Using discourse analysis methods, sequential modeling, and text classification, we will explore the potential of modeling conversational context and trajectory to select the most appropriate strategies for collaborative facilitation.
Acknowledgments

Thanks to
Barry, Faith, Kate
Carolyn, Vincent, Alex, Diane
Dave, Colin, Jack, Oscar, Tim, Nicole
Sherice, Chuck, Lauren, Gaowei, Kate, Stefani, Sandy
Greg, Iris, Hyeju, Ryan, and Elijah!
## Contents

1 Introduction  

2 Theoretical Framework  
   2.1 Supporting Effective Collaborative Discussion  
   2.2 Dynamic Script-Based Support with Conversational Agents  
   2.3 Dynamic Support for Collaborative Learning with Academically Productive Talk  
   2.4 Predicting Conversational Outcomes  

3 An Architecture for Dynamic Conversational Facilitation  
   3.1 Dynamic Support for Collaboration  
   3.2 The Bazaar Conversational Agent Architecture  
      3.2.1 Events and Components  
      3.2.2 Output Coordinator: Prioritizing Proposed Actions  
   3.3 Bazaar in the Wild  

4 Accountable Talk Agents  
   4.1 Detecting Academically Productive Talk Candidates  
   4.2 Revoicing Facilitation  
   4.3 Agree-Disagree Facilitation  
   4.4 APT Feedback  

5 Automated Accountable Talk Facilitation in Action  
   5.1 Experimental Paradigm  
   5.2 Assessment  
   5.3 Study 1: 9th Grade Diffusion Lab  
   5.4 Study 2: 9th Grade Genetics  
      5.4.1 Participants  
      5.4.2 Experimental Manipulation  
      5.4.3 Study Procedure  
   5.5 Study 3: Freshman Engineering Design  
      5.5.1 Participants  
      5.5.2 Experimental Procedure  
      5.5.3 Results  
   5.6 Study 4: Freshman Honors Chemistry  
      5.6.1 Participants  
      5.6.2 Experimental Manipulation  
      5.6.3 Experimental Procedure  
      5.6.4 Results  
   5.7 Discussion  
   5.8 Conclusions
6 Proposed Work

6.1 Proposed Study: Full-Spectrum Facilitation ...................................... 29
6.2 Proposed Study: What Would Teachers Do? ........................................ 30
6.3 Proposed Work: Modeling Conversational State to Predict Facilitation Opportunities ........................................ 31
    6.3.1 Active Learning Annotation ....................................................... 31
    6.3.2 Contrastive Hidden Markov Models ............................................. 31
    6.3.3 Silent Signals for Facilitation .................................................... 32
    6.3.4 Evaluation ............................................................................. 32
6.4 Proposed Study: Tying it Up With a Bow ............................................ 32
6.5 Power Tools and Generalization ......................................................... 33

7 Summary and Timeline

7.1 Summary of Contributions ................................................................. 35
7.2 Timeline ....................................................................................... 35

Bibliography ................................................................................... 37
List of Figures

3.1  The chat interface, showing a collaborative facilitation interaction. ........................................ 10
3.2  Overview of the Bazaar event flow. ......................................................................................... 12
3.3  The Output Coordinator’s orchestration of proposed actions. .................................................. 13
4.1  Triggering Agree-Disagree facilitation. .................................................................................... 17
5.1  Punnett Square problems from 9th grade genetics unit. .......................................................... 22
## List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Accountable Talk Moves</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>Component Configurations used in current studies</td>
<td>14</td>
</tr>
<tr>
<td>4.1</td>
<td>Examples of Revoicing Behaviors</td>
<td>16</td>
</tr>
<tr>
<td>4.2</td>
<td>Revoicing example in a 9th grade biology lesson from a Genetics unit</td>
<td>16</td>
</tr>
<tr>
<td>4.3</td>
<td>Agree-Disagree example in a 9th grade biology lesson from a Genetics unit</td>
<td>17</td>
</tr>
<tr>
<td>4.4</td>
<td>Agree-Disagree example in an undergraduate Chemistry unit on intermolecular forces</td>
<td>18</td>
</tr>
<tr>
<td>4.5</td>
<td>APT Feedback example in a high school Biology unit on diffusion</td>
<td>18</td>
</tr>
<tr>
<td>5.1</td>
<td>Example of Feedback and Revoicing in the Dyke et al. study.</td>
<td>21</td>
</tr>
<tr>
<td>5.2</td>
<td>Summary of Results Across Studies</td>
<td>27</td>
</tr>
<tr>
<td>6.1</td>
<td>Additional Features</td>
<td>30</td>
</tr>
<tr>
<td>7.1</td>
<td>Proposed Timeline</td>
<td>36</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The need for quality educational experiences online and at scale, has become painfully evident amidst the recent boom in massive and open online courses. While instructor time is a scarce commodity, students are plentiful. Although this disparity is deeply underlined by such lessons of mass instruction, teachers in traditional classrooms also struggle to support rich discussions between students, against the pressures of standardized testing and declining teacher-student ratios. With this imbalance in mind, one important contribution the field of intelligent support for group learning can make is to develop technologies to structure and support interactions between students. Effective collaborative learning experiences are known to provide many benefits to learners in terms of cognitive, metacognitive, and social impact [41, 71, 84]. In order to support collaborative learning at any scale, affordances must be provided to support high-quality interactions in the absence of human facilitators. Effective, automated support for such interactions is the key to this facilitation. However, such automation relies on first being able to model and assess the factors that predict learning.

We ground this proposal in a paradigm for dynamic support for group learning. This paradigm, featuring tutorial dialogue agents, has proven effective for improving interaction and learning in a series of online group learning studies [5, 16, 17, 44, 46, 47]. Such supports are dynamic in that they can be triggered through real time analysis of the collaborative discussion as it unfolds [1, 26, 44, 46]. The decision-making process involves identifying triggers in the ongoing collaboration in real time, and then launching a specific supportive behavior at the appropriate time in response to those triggers. A specific pairing of a trigger method and an intervention can be thought of as a strategy for facilitation. In our prior work, each study has explored a single strategy that was meant to behave dynamically, according to the same context-sensitive rules for all student groups. In this proposal, we explore how the available support strategies might be adapted depending upon the characteristics of the student population, and of an individual group’s dynamic. In particular, we build on prior work in triggering support based on real time analysis of collaborative discourse and work towards a new paradigm for dynamic support.

The analysis of our completed work integrates findings from a series of completed studies across content domains (biology, chemistry, engineering design), grade levels (high school, undergraduate), and facilitation strategies. In each study, each experimental condition makes use of only one strategy. As we observe the pattern of results across studies, where the studies differ in domain and grade level, we see that the ranking among strategies in terms of the relative effectiveness of alternative strategies differs depending on the student population and learning task. We also observe a characteristic pattern in the interaction between students within successful conditions that can be detected with high reliability through automated collaborative process analysis. This series of studies, along with the automated process analysis technique, provide an initial empirical foundation for the development of a more agile approach to supporting group learning. By adapting the choice and timing of facilitation strategies in response to both initial context and the automated assessment of student interaction patterns, we may be able to provide better support when and how it’s needed, and get out of the way when it isn’t.

Prior work has shown that models learned from sequences of discourse acts can predict learning and social outcomes, even from just the early stages of an ongoing discussion [73, 83], and that automatable
representations of a conversation’s state hold benefit for both human and automated support [11]. We propose using sequential annotations and textual features to model evolving facilitated discussions. In addition to anticipating the outcomes of a collaborative activity, such models may be used to dynamically classify ongoing conversations as having characteristics which suggest one or more appropriate facilitation strategies.

As the primary contribution of this proposed work, we plan to develop and operationalize a framework for automated, context-responsive facilitation. First, we shall examine the suitability of additional Academically Productive Talk facilitation strategies, and examine how this varies with student population and factors of group composition. Further, through a set of analyses on the discourse of these and earlier transcripts, we seek to explain the mechanisms by which these facilitation moves affect conversational and learning outcomes in each context. We will also employ seasoned teachers and facilitators to annotate these transcripts with their judgement of where the most opportune moments for facilitation occur, to validate our approach and provide fodder for a new, learned model for facilitation triggering.

In Chapter 2 we first describe a theoretical foundation from prior work in the literature on computer supported collaborative learning, tutorial dialogue agents, and classroom discourse. In Chapter 3 we describe our technical approach, built around Bazaar, a publicly available architecture for dynamic support for collaborative learning. Chapter 4 details a set of facilitation techniques, implemented in Bazaar and drawn from prior work in classroom discourse, which we investigate through a set of completed studies described in Chapter 5. By integrating the results presented in the individual studies, we motivate a research agenda for future work in the area of intelligent support for group learning. In Chapter 6 we describe the work yet to be completed. Chapter 7 offers a brisk timeline for the proposed works’ completion, and summarizes the contributions of this thesis.
Chapter 2

Theoretical Framework

The theoretical foundation for the work discussed in this proposal comes from three areas. We begin with literature from the Computer Supported Collaborative Learning (CSCL) community. Here we draw insights into types of conversational interactions that are associated with learning in groups and typical static technology for increasing the prevalence of those types of interactions, and thereby increasing learning. Next we review more recent work from the CSCL community where dynamic forms of support for group learning have been developed and demonstrated to be advantageous over more typical static forms of support. We review the classroom discussion facilitation literature that motivates the set of dynamic support strategies we evaluate in this paper. Finally, we propose that these strategies can serve as building blocks for a new form of dynamic, domain-independent “agile” support for group learning, rooted in the best practices of expert classroom facilitators.

2.1 Supporting Effective Collaborative Discussion

The field of Computer Supported Collaborative Learning (CSCL) has a rich history extending for nearly two decades, covering a broad spectrum of research related to learning in groups, especially in computer mediated environments. A detailed history is beyond the scope of this article, but interested readers can refer to Stahl’s well known history of the field [78] and other foundational work [20]. An important technological goal of work in the field of CSCL is to develop environments with affordances that support effective group learning. The foundation for this work comes from insight into the patterns of conversational interactions that are valuable for learning. A series of studies in the computer-supported collaborative learning field demonstrate the pedagogical value of social interaction from a cognitive perspective, showing that interventions that intensify argumentative knowledge construction, in support of group knowledge integration and consensus building, enhances the development of multi-perspective knowledge [87, 88].

Despite differences in orientation between various flavors of learning science, some of the conversational behaviors that have been identified as valuable are very similar across sub-communities. Such frameworks for characterizing conversational behaviors share two aspects:

- The requirement for reasoning to be explicitly displayed in some form.
- The preference for connections to be made between the perspective of one student and that of another.

We base our work on such characterizations of productive discussion behavior. Frameworks for analysis of group knowledge building that favor subtly different formulations of these behaviors are plentiful. In particular, these include transactivity [12, 81, 87], Inter-subjective Meaning Making [80], and Productive Agency. Similar arguments for the significance of these kinds of behaviors can be made from multiple perspectives, Schwartz [72] from a Sociocultural perspective and De Lisi and Golbeck [19] from a Piagetian Cognitivist perspective. The idea of transactivity comes originally from a Piagetian framework. The process is explained similarly to how we describe the production of transactive contributions. In both cases, mental models are explicitly articulated, shared, mutually examined, and possibly integrated.
The most popular formalization of the construct of transactivity \cite{13} includes 18 types of transactive moves. These characterize conversational turns that can be considered explicit reasoning displays that connects to some previously articulated reasoning display. Within this schema, transacts have been divided along multiple different dimensions, which we will draw from later to motivate our series of experimental studies. One important dimension represents whether the transact might be self-oriented (the contribution operates on the speaker’s own reasoning) or other-oriented (the contribution operates on the reasoning of a partner) \cite{13, 81}. Another important dimension is whether the contribution represents the original idea as stated, or transforms it into something new. Another dimension is whether the contribution is consensus-oriented (extending or affirming earlier ideas) or conflict-oriented (challenging established ideas).

In order to support the growth of student discussion skills, it is necessary to design environments with affordances that encourage transactive behaviors and other valuable learning behaviors. The most popular approach to providing such affordances in the past decade has been that of script-based collaboration \cite{22, 42, 43}. A script is a schema for offering scaffolding for collaboration. Some typical forms of scripts come in the form of instructions that structure a collaborative task into phases, or structured interfaces that reify certain types of contributions to the collaboration. Such scripts are typically implemented statically, providing the same support in all cases. A script may describe any of a wide range of features of collaborative activities, including its tasks, timing, the distribution of roles, and the methods and patterns of interaction between the participants. Static scripts do not behave differently depending on what is happening in the collaboration per se. Instead, they operate according to choices that are made ahead of time and generally held constant within conditions in an experimental study.

Collaborative scripts can be described as operating on either the macro- or micro- level \cite{21}. Macroscripts are pedagogical models that describe coarse-grained features of a collaborative setting. They can sequence and structure each phase of a group’s activities to foster learning and social interaction. Microscripts, in contrast, are models of dialogue and argumentation that are embedded in the environment, and are intended to be adopted and progressively internalized by the participants. Scripts (at either scope) can be more or less coercive, from strict “follow-me” style prompts to subtle suggestions of behavior implicit in the activity’s structure. Stricter scripts can work to reduce the gap between expected and observed student behavior, producing a more uniform appearance of discussion. However, they run the risk of over-scripting \cite{22}, where the application of inappropriate or unneeded supports have a detrimental effect on collaboration and learning.

### 2.2 Dynamic Script-Based Support with Conversational Agents

The early non-adaptive scripting approaches described above can sometimes result in both over-scripting and in interference between multiple scripts \cite{88}, both of which have been shown to be detrimental to student performance. More dynamic approaches can trigger scripted support in response to the automatic analysis of participant activity \cite{27, 56, 60, 69, 74}. This sort of analysis can occur at a macro-level, following the state of the activity as a whole, or it can be based on the micro-level classification of individual user contributions. Some prior work on adaptive support for collaborative learning used hint-based support for individual learning with technology to support peer tutoring interactions \cite{23}. Other prior work on dynamic conversational agent based support built on a long history of work using tutorial dialogue agents to support individual learning with technology \cite{31, 68, 89, 91}.

The collaborative tutoring agents described by Kumar and colleagues \cite{45, 46} were among the first to implement dynamic scripting in a CSCL environment. In that work, the role of the support was to increase conceptual depth of discussions by occasionally engaging students in directed lines of reasoning called Knowledge Construction Dialogues (KCDs) \cite{70} that lead students step by step to construct their understanding of a concept and how it applies to the collaborative problem solving context. These encounters were triggered in the midst of collaborative discussions by detection that students were discussing an issue that is associated with one of the pre-authored interactive directed lines of reasoning. Thus, these interventions had the ability to be administered when appropriate given the discussion, rather than being triggered in a one-size-fits-all fashion. In an initial evaluation \cite{46}, this form of dynamic support was associated with higher learning gains than a control condition where students had access to the same lines of reasoning.
but in a static form. In a subsequent study, students were found to gain significantly more if they had the option to choose whether or not to participate in the directed line of reasoning when it was triggered \[16\]. Scripting such as this offers the potential for minimal interventions to be used more precisely and to greater effect, with greater likelihood of students internalizing the support’s intended interaction patterns. Further, the benefits of fading support over time \[50\] could be more fully realized, as the frequency of intervention could be tuned to the students’ demonstrated competence.

A major limitation of the specific form of interactive support provided by KCDs is that by their very nature they are content specific. Thus, for every new concept, a separate authoring effort was necessary, which limits the scalability of the approach.

### 2.3 Dynamic Support for Collaborative Learning with Academically Productive Talk

A promising direction for addressing the issue raised above related to content specificity is to draw inspiration from the classroom discourse literature, where content independent strategies for eliciting valuable interaction between students have been developed and tested. One notable framework for such elicitation is Academically Productive Talk (APT) \[57\]. APT, also called “Accountable Talk,” is a classroom discussion facilitation approach that has grown out of instructional theories that emphasize the importance of social interaction in the development of mental processes. In particular, it values those interactions that engage students in transactive exchanges. Drawing on over 15 years of observation and study, Michaels, O’Connor and Resnick propose a number of core facilitation “moves”, displayed in Table \[5.1\]. These serve as tools that teachers can employ in order to encourage the development of productive classroom discussions – in particular, discussions in which students make their reasoning public, listen deeply and critically to one another’s contributions, and then interact with them transactively. These facilitation techniques serve both to guide the discussion and to model productive collaborative conversational behavior. Academically Productive Talk thus serves as a collaborative script that can shape group discussion. Students who are more advanced may internalize this script, and thus require less external facilitation.

The set of APT moves is presented as a domain-independent collaboration script. However, there are other scripts for discussion and argumentation that are particular to the practices of a given field. Prior work has suggested that modeling such practices in classroom discussion is beneficial. The general practices of Academically Productive Talk can complement and reinforce such domain-specific scripts, and offer greater opportunity for transference to new domains \[64\].

Our recent efforts have developed intelligent conversational agent facilitators whose behavior is not content or subject-specific, but rather draws from this literature on facilitation strategies \[1, 18, 26\]. The design of such support is consistent with the literature on facilitation of collaborative learning groups \[35\], and leverages the large body of work that has shown that APT facilitation behaviors are beneficial for learning with understanding \[4, 14, 15, 65, 66, 82, 86\].

In earlier published studies where teachers used approaches like Academically Productive Talk, students have shown steep changes in achievement on standardized math scores, transfer to reading test scores, and retention of transfer for up to 3 years \[4, 14, 15, 65, 66, 82, 86\]. These successes in the classroom discourse literature offer hope that these facilitation strategies could be used to design effective support for collaborative learning, a concept we refer to as APT agents. However, none of these earlier studies have explored the question of what the preconditions for successful use of specific APT moves might be, or what kinds of learners would benefit most from which facilitation moves. Nevertheless, this kind of detailed insight is needed if these moves are to be used to their maximum benefit as support for collaborative learning.

The set of Academically Productive Talk moves includes paraphrasing a student statement: “So let me see if I’ve got your thinking right. You’re saying XXX?”, which encourages students to reformulate or transform the articulation of their reasoning in order to clarify their meaning. Another move involves asking students to apply their own reasoning to someone else’s reasoning: “Do you agree or disagree, and why?”, which may stimulate sociocognitive conflict, otherwise known as conflict-oriented consensus building. As we have illustrated in Table 1, these core moves can be characterized in terms of the type of
### Table 2.1: Accountable Talk Moves

<table>
<thead>
<tr>
<th>Example Teacher Utterance</th>
<th>APT Move</th>
<th>Self vs. Other</th>
<th>Represent vs. Transform</th>
<th>Consensus vs. Conflict</th>
<th>Accountable Talk Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you say more about that?</td>
<td>Say More</td>
<td>S</td>
<td>R</td>
<td>C</td>
<td>Step 1: help individual students share and clarify their own thinking</td>
</tr>
<tr>
<td>Let me see if I understand you. Are you saying they were all adopted?</td>
<td>Revoice</td>
<td>S</td>
<td>R</td>
<td>C</td>
<td>Step 1: help individual students share and clarify their own thinking</td>
</tr>
<tr>
<td>What’s your evidence? How did you arrive at that conclusion?</td>
<td>Press for Reasoning</td>
<td>S</td>
<td>T</td>
<td>C</td>
<td>Step 3: help students deepen their own reasoning</td>
</tr>
<tr>
<td>If capital G is dominant, wouldn’t all babies be orange?</td>
<td>Challenge</td>
<td>S</td>
<td>T</td>
<td>X</td>
<td>Step 3: help students deepen their own reasoning</td>
</tr>
<tr>
<td>Can you repeat what she said?</td>
<td>Restate</td>
<td>O</td>
<td>R</td>
<td>C</td>
<td>Step 2: help students listen carefully to one another</td>
</tr>
<tr>
<td>Who can explain what Aisha means when she says that?</td>
<td>Explain Other</td>
<td>O</td>
<td>R</td>
<td>C</td>
<td>Step 4: help students think with others</td>
</tr>
<tr>
<td>Help him out Stephen. Can you add to what he said?</td>
<td>Add More</td>
<td>O</td>
<td>T</td>
<td>C</td>
<td>Step 4: help students think with others</td>
</tr>
<tr>
<td>Kelly, are they right? Do you agree or disagree with what they said, and why?</td>
<td>Agree Disagree</td>
<td>O</td>
<td>T</td>
<td>X</td>
<td>Step 4: help students think with others</td>
</tr>
</tbody>
</table>

Transactive behavior they might elicit from students along the three dimensions we introduced above. It is important to note that across these dimensions, these types of transacts can be seen as having a logical ordering which might then apply to the corresponding APT facilitation moves as well. For example, one must understand one’s own reasoning before one can hope to understand another person’s reasoning, thus self-oriented transacts could be seen as less demanding than other-oriented ones. Furthermore, one must understand reasoning as stated before one can transform or extend that reasoning, thus representational transacts might be seen as less demanding than transformational ones. Reasoning must be understood before it can be rightly challenged, thus, it would be possible to argue that conflict oriented consensus building requires more than consensus oriented transactive behavior. Some prior work has attempted to tease apart differential meditational effects of transacts from these various categories [7]. Building upon this foundation, it is reasonable to hypothesize that the specific APT move that would be helpful to students would depend upon the student’s specific capabilities or the difficulty of the material being discussed.

Michaels and O’Connor describe Academically Productive Talk “as moves that support externalization of student thinking, and the work that teachers do to make sure that the thinking is intelligible to everyone.
else, and that everyone is listening to and taking seriously the thinking of one's peers.” They argue that because the conventional ordering of transactive moves is based largely on work with dyads, it undervalues some aspects of productive group discussion, including externalization (Say More and Revoice) and listening to others (Restate), the latter of which is not typically coded in transactive analysis. Because these are not internalized practices for many students and classrooms, it can be necessary to support even these basic needs. Transactive analysis of dialogues has shown that students benefit most from exposure to discourse at a slightly higher transactive level than their own [81]. The ordering of APT goals, closely related to that of transacts, may well benefit from the same sort of targeting. This is suggested by the APT literature, but not explicitly verified.

The framing of APT also diverges from the transactivity literature by focusing on what teachers do to support and model transactive and self-explanation behaviors, rather than on the student behavior by itself. On its own, the transactivity literature doesn’t give us concrete guidance on when and how to manage learners and actively facilitate learning. To that end, APT leverages observations of teachers skilled in managing group learning to outline these moves as a set of replicable interventions. Tools designed to support productive discussion might benefit from this approach. By modeling the responses of teachers, such tools might be able to automatically detect opportunities for facilitation, and respond to them in a manner consistent with observed practice.

### 2.4 Predicting Conversational Outcomes

Linguistic analysis methods for studying both individual learners and small groups [39] have been used to assess cognitive and meta cognitive knowledge [34], critical thinking, knowledge construction [52] and consensus building techniques [49]. In many cases [24, 83], methods for automatically labeling these features are developed hand-in-hand with their application to a prediction task. Analysis applied to course message boards has shown it is possible to detect unresolved questions [40] in asynchronous discussions, and that patterns of interaction and participation can be used to predict final learning outcomes [67].

In the context of a single-user conversational tutor, a set of conversational features, including measures of the quality and content of student answers as derived from Latent Semantic Analysis [48], have been successfully applied to predict the moment-to-moment affect of the learner [24].

In intelligent tutoring systems with a conversational component, automated analysis methods may be employed as formative assessments, predicting student learning or collaborative performance. These predictions can be used to inform a tutor’s interventions during future learning experiences, or to provide moment-by-moment facilitation in response to continuous assessment [3]. Recent work has demonstrated the power of data mining for building moment-to-moment models of student learning [9], although as this work was situated in a non-conversational tutoring system, it did not leverage linguistic features to anticipate learning. Fully automated coding and modeling methods have been used to successfully predict the outcome of a facilitated civil-dispute negotiation [83]. Models of conversational trajectory have also been developed as a source of feedback for learners and their human instructors, using a set of features describing conversational attributes derived from per-turn coding of a conversation [10, 11]. In that work, each coded move contributes to one of four underlying conversational dimensions (conformity, creativity, elaboration, and initiative), allowing concrete quantitative measures to power a qualitative analysis of group state.

Hidden Markov Models [63] trained on sequences of student-selected sentence-opener moves have been used to classify and describe groups of collaborative learners as more or less productive [76, 77]. HMMs have also been applied to surveys of participant emotion, to draw inferences about underlying affective or cognitive state [25]. However, such work has relied on participants selecting their next move or observed state from a limited set of options. More recent work has used n-grams or stretchy patterns [30] over discourse act labels to model local conversational structure and predict group task success [55]. Although this body of work illustrates the potential of sequential models for understanding student state, their suitability as a method for assessing individuals within an unconstrained multiparty discourse has not been fully explored.
Chapter 3

An Architecture for Dynamic Conversational Facilitation

The publicly available Bazaar architecture enables easy integration of a wide variety of discussion facilitation behaviors that has enabled the set of experimental studies we describe in the next section. We begin this section by describing from a user perspective one integrated environment where Bazaar provides collaboration support to distributed groups of learners collaborating synchronously. Next we describe the inner workings of the architecture and how it enables effective coordination of supportive facilitation behaviors. We then discuss how we have used this resource to implement the facilitation behaviors we evaluate in our experimental studies.

3.1 Dynamic Support for Collaboration

The Bazaar architecture [2] has been used in a variety of studies [1, 13, 20, 37] to implement supportive interventions involving conversational chat agents that participate as facilitators in collaborative learning tasks. The architecture has been successfully integrated with a variety of collaborative environments. These include Moodle chat activities, generic web-based chatrooms, and specialized collaborative environments with shared workspaces [36, 61]. Figure 3.1 displays an integration between Bazaar and the ConcertChat [61] synchronous chat collaboration environment, which was used for the studies described in Chapter 5. Because the Bazaar architecture enables quick development of supportive interventions, one can efficiently proceed from a concept for a new support behavior to a fully functional collaboration environment. In Figure 3.1 the panel on the right hand side of the interface is a chat panel where students interact with one another through synchronous chat. The turns labelled as ‘Tutor’ are turns that come from the intelligent conversational agent providing facilitation moves in the conversation. In this example we see the agent performing a Revoicing move. On the left is a shared white board where either the agent or the students can insert images that are then visible to the whole group. In this case, the image displays a cell model that the students were meant to discuss in the Diffusion Lab. The relative size of the chat panel and the white board can be adjusted by clicking in between the two panels and dragging in one direction or the other.

3.2 The Bazaar Conversational Agent Architecture

Bazaar is a modular framework for designing multi-party collaborative agents that builds upon the earlier Basilica architecture [26, 45]. Like Basilica, in addition to its core architecture, Bazaar plays host to a library of reusable behavioral components that each trigger a simple form of support. More complex supportive interventions are constructed by integrating multiple simpler behaviors. For example, in the Dyke et al. [26] study, in the condition with both Revoicing and Feedback, the agent needed to coordinate
the macro-level prompts with the micro-level prompts from both the Revoicing and Feedback strategies. Both the agent’s overall composition and the configuration of each component are specified in plaintext properties files, offering a glimpse at the sort of low-overhead flexibility for authoring, content, and deployment championed by recent work \[42\]. Bazaar and its predecessor are event-driven systems in which independent behavioral components receive, filter, and respond to user, environment, and system-generated events, and present the unified output of these components to the user. Bazaar improves on the Basilica architecture by integrating the orchestration of otherwise competing or conflicting agent behaviors, by simplifying the relationships between components, and by offering an extensible mechanism selecting proposed agent actions. The issue of potential clash between macro-level support and micro-level support is especially important, as we have observed that experiencing these clashes is distracting and confusing for students \[37\]. Thus, it is important to note that coordination between simple support behaviors is necessary even when only one APT facilitation strategy is being used. Figure 3.2 illustrates a typical Bazaar configuration where events triggered by student contributions in the chat or whiteboard are aggregated in the Input Coordinator. Unlike Basilica, event processing in Bazaar is divided into two distinct phases. Preprocessor components analyse the event stream in search of triggers for supportive interventions. Two examples are shown in Figure 3.2 including the Revoicable Annotator, which looks for student turns that could be revoiced by the agent, and the Participation Counter, which keeps track of how many utterances each student has contributed recently. These preprocessed events are relayed to a set of Reactors components. Depending on the active agent strategies, under specified circumstances, these Reactors will propose tutor actions in response to these events. The Output Coordinator, described in the next section, then makes decisions about sequencing and timing and thus manages the coordination of potentially clashing interventions. Thus, the Output Coordinator controls when the prompts or other behaviors associated with a triggered strategy are presented to the students.

The Output Coordinator houses Bazaar’s primary architectural improvement. In an agent able to offer multiple dynamic behaviors, more than one support strategy may be simultaneously appropriate. Bazaar’s predecessors sometimes suffered from clashes between behaviors in cases where multiple were triggered simultaneously. It is important to note that the interference of multiple supports caused by these clashes could invalidate the benefit of any of them, to the detriment of the learner \[32, 88\]. It is important to note that participants in a collaborative session, including the facilitator, are not simply focused on the task they are involved in numerous simultaneous processes including social bonding, idea formation, argumentation, time management, and off-task activity. Managing an APT discussion poses
additional challenges. While the kind of in-depth discussion that APT elicits is valuable for learning, it takes time. Facilitators must always keep time constraints in mind in order to achieve an appropriate balance of breadth and depth within and across topics as well as in parcelling out attention to different students.

As we have alluded to, we observed problems with time management in an earlier prototype implementation of an APT agent implemented using Basilica [37] that manifested as clashes between the macro and micro scripting behaviors triggered during the study. As a technical solution to this multi-policy management problem, Bazaar draws on and extends the "concurrent mode" approach described by Lison [50]. In Lison’s work, the author adds a "soft" constraint on new proposals by increasing the relative weight of those from the same source as recent actions, preferring that source as a "focus of attention" for as long as it had new actions to propose. Proposals with a great enough activation weight (or priority) from different sources can outweigh this preference, allowing flexible yet consistent responses in the face of noisy input or multiple valid states. Evaluation in a simulated human-robot learning task showed that this "soft" control method performed better than using a hierarchical finite-state controller to select the next source of action. We apply this approach in Bazaar, allowing recent actions to influence the priority of new proposals, and extend it, allowing recent actions to promote or suppress proposals from any source.

In the sections that follow, we describe Bazaar’s event flow in more detail, and the way in which it affords flexible orchestration between multiple behavioral components. This orchestration is key to providing agile, responsive conversational supports. It also underpins Bazaar’s role as a rapid research platform.

3.2.1 Events and Components

In Bazaar, an Event is an object representing something interesting that has happened in the world of the agent. Some Events come from the environment and map to the actions of participants, like a user entering a chat room, or an incoming user message; these may be annotated by Preprocessor components to reflect a rich understanding of the Event. New Events can also result from the analysis of other Events, or represent awareness of system state. Events such as these are used to launch phases of macro-scripts, or to initiate dynamic support. Bazaar components can generate and respond to arbitrary author-defined Events, thus it is not possible to provide a comprehensive list. The default Event classes handled by the core Bazaar components include Message (a chat message is sent by a student), Presence (a student enters or leaves the chat room), Whiteboard (a student manipulates an object in the shared whiteboard), Dormancy (a student or group has been idle for a certain amount of time), Launch (author-specified conditions for beginning a macro-script have been met), and Step Done (a stage in a macro-script step has been completed).

Components in Bazaar represent a modular representation of related behavior and state-knowledge, corresponding to all or part of a single method of scripting or support. Components respond to those Events they consider relevant. Bazaar defines a two-step event-processing flow, dividing components’ event-processing responsibility into Preprocessor and Reactor roles. While some components may act in both roles, this two-stage processing is still enforced. When a new Event is received by the system, all Preprocessor components that have registered for a particular Event class are given the opportunity to respond to it. They may respond by generating new Events (perhaps to indicate a shift in the conversation’s focus) or by adding information to the original Event. Events are subsequently delivered to those Reactor components which are registered for these Events’ classes. Reactors have the opportunity to respond to preprocessed Events (to dynamically enact sub-scripts or supports) by proposing actions to the Output Coordinator.

3.2.2 Output Coordinator: Prioritizing Proposed Actions

As mentioned above, the Output Coordinator is needed to avoid clashes between multiple proposals that may have been triggered within the same period of time. Most commonly, clashes occur between proposals related to macro level support and proposals related to micro level support. Figure 3.3 illustrates an example proposal flow within the Output Coordinator. Proposals for agent action, received from the
Reactor components, are queued by Bazaar’s Output Coordinator. When a Reactor creates a Proposal, it is assigned a timed window of relevance, and a priority (between 0 and 1). Periodically, the Output Coordinator will re-evaluate the priority of each remaining Proposal (by taking hints from recently enacted Proposals), rejecting those that have expired, and accepting and enacting the Event with the highest priority. A previously-accepted agent action can leave a lingering presence with the Output Coordinator, a Proposal Advisor, which can re-weight the priority of (or entirely suppress) incoming Proposals until its influence expires. Each action Proposal is constructed with a timeout-window after which it is no longer relevant - if a queued Proposal has not been accepted when its timeout expires, it is removed from the queue. When a message is accepted or rejected, a callback method (which may be defined at the time of Proposal creation) is invoked, allowing the proposing Component to update its state accordingly.

Bazaar provides methods for creating Proposals with Proposal Advisors for common use cases. These include sending simple single turn messages, or interventions that involve sequences of messages and that suppress all subsequent Proposals (or those from a particular set of source components) for a given amount of time or until the sequence of associated behaviors is complete (to allow an opportunity for student follow-up, for example). In most cases, employing these pre-defined advisors is sufficient to author a smooth and natural agent experience. Bazaar also supports more advanced proposal-management techniques, such as affording a Proposal the ability to re-evaluate its own importance in light of subsequent Events.

By allowing Proposals to establish constraints on near-future Events in a general way, conversational agents authored in Bazaar can be responsive to changes in both student behavior, and in the behaviors enacted by the agents’ behavioral components. As support behaviors re-evaluate their own relevance, the agent thus has the potential to effectively change strategies dynamically, based on whether the current
strategy is having the desired effect. Authors of Bazaar agents can specify these to suit their experimental, pedagogical, and practical needs. In particular, the rigidity of timing with which macro-scripted elements are executed can be adjusted along the spectrum between replicability and internal experimental validity, and natural, external conversational validity. Table 3.1 details the Proposal and Advisor configurations for components used in the studies described in Chapter 5. The components themselves are described in detail in Chapter 4.

### 3.3 Bazaar in the Wild

Bazaar agents have been used in several recent research studies [37, 38]. A series of such studies is discussed in detail in Chapter 5. In many cases, the agents have been deployed remotely and at scale, to dynamically-assembled collaborative groups at distant sites. The architecture’s modularity and configurability ease the rapid development and coordination of multiple agent behaviors. Bazaar has been used as a teaching tool in several university courses and learning-science workshops, where students successfully developed and extended conversational agents with minimal incoming experience [2]. The ease of authoring and integrating new agents and behaviors underlines the strength of Bazaar’s contribution as a research platform.
Table 3.1: Component Configurations used in current studies

<table>
<thead>
<tr>
<th>Bazaar Component</th>
<th>Behavior Intent</th>
<th>Proposal Priority</th>
<th>Proposal Timeout</th>
<th>Advisor Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timed Script</td>
<td>Provide consistent time for each section across groups, allow time for reading</td>
<td>High</td>
<td>60 seconds</td>
<td>Block all tutor actions for a time proportional to the length of the displayed prompt.</td>
</tr>
<tr>
<td>Social Support</td>
<td>Offer immediate responses to social cues</td>
<td>Low</td>
<td>3 seconds</td>
<td>Block all tutor actions for 5 seconds.</td>
</tr>
<tr>
<td>APT Feedback</td>
<td>Give immediate feedback on student APT behaviors</td>
<td>High</td>
<td>3 seconds</td>
<td>Block all other tutor actions for 5 seconds, block other APT moves for 20 seconds</td>
</tr>
<tr>
<td>Revoicing</td>
<td>Highlight and clarify student-generated concepts</td>
<td>Medium (proportional to candidate similarity)</td>
<td>15 seconds</td>
<td>Block all other tutor actions for 10 seconds, block other APT moves for a further 45 seconds</td>
</tr>
<tr>
<td>Agree Disagree</td>
<td>Support discussion of student-generated concepts</td>
<td>Medium (proportional to candidate similarity)</td>
<td>15 seconds</td>
<td>Check for student followup before acting. Prioritize agree-disagree tutor followup prompts. Block other tutor actions for 10 seconds, block other APT moves for a further 45 seconds</td>
</tr>
</tbody>
</table>
Chapter 4

Accountable Talk Agents

Three different interventions based on Academically Productive Talk strategies are evaluated in the series of studies discussed in Chapter 5. These strategies Revoicing elicits Self-Oriented, Transformational, Consensus Oriented transacts. Agree-Disagree elicits Other-Oriented, Representational, Conflict-Oriented transacts. Finally, APT Feedback is designed to offer non-specific encouragement for students to engage in APT related behaviors. In this chapter, we describe the implementation of these facilitation strategies in detail.

4.1 Detecting Academically Productive Talk Candidates

The two APT interventions implemented for the studies reported in this paper required the detection of task-relevant conceptual assertions. For example, attempts at articulation of task-relevant assertions could be the focus of a reformulation elicited by a Revoice facilitation move or the idea that a student agrees or disagrees with in response to an Agree/Disagree move.

In order to identify task-relevant conceptual assertions, we worked with domain experts and instructors to develop a gold standard list of statements that captured important concepts and misconceptions for the unit of study. Such statements were drawn from both the experts' knowledge and expectations and from transcripts of an unsupported dry-run of the task. We adopt a “bag of synonyms” cosine similarity measure \[29, 58\], which essentially measures overlap in word usage. Student assertions which are within a certain threshold of similarity to the gold statements are identified as revoicable or agree-disagree candidates that could be evaluated by the group. Both the Revoicing and Agree-Disagree supports described employ use the same detection method (implemented as a Bazaar Pre-Processor component), although with a looser similarity threshold in the latter case.

4.2 Revoicing Facilitation

One of the forms of support evaluated in this paper is a Bazaar agent that performs the APT Revoicing move. The agent compares student input against a list of correct statements drawn from the data collected in pilot runs of the studies. If an entry in this list could be interpreted as a paraphrase of the student’s input using the method described above, it is offered by the agent as a “revoicing” to the students. The same statement was never offered more than once in the same session as a revoicing. When student statements were not close enough to match the revoicing list but contained the first mention of important lesson concepts (like “test strip” or “molecule size”), the agent would ask the student or a peer to expand or restate their contribution. Examples are given in Table 4.1.

An example from a unit of 9th grade biology on Genetics, which was the context for Study 2 discussed below, is displayed in Table 4.2. Here all of the student turns that are detected to be revoicable are marked with italics. The Tutor’s revoicing is marked in bold. Note that while two turns were detected as revoicable by the system, a revoicing was only triggered once because of the constraint that the same
Table 4.1: Examples of Revoicing Behaviors

<table>
<thead>
<tr>
<th>Student Contribution</th>
<th>Revoicing Facilitation Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>basically the glucose will get inside</td>
<td>Maybe you could state that as “the cell membrane is permeable to glucose.”</td>
</tr>
<tr>
<td>it changed because the tube absorbed the iodine,</td>
<td>So are you saying “the molecules diffused through the membrane?”</td>
</tr>
<tr>
<td>I predict that if the holes in the plastic are large enough, the glucose will go into the water solution..</td>
<td>Maybe you could state that as “both water and glucose molecules are able to move between the two environments.”</td>
</tr>
</tbody>
</table>

concept won’t be revoiced more than once in the same conversation. What we see in this example is that the tutor’s revoicing of Student1 created the opportunity for that idea to be the focus of reformulation and clarification, as shown by Student2’s followup.

Table 4.2: Revoicing example in a 9th grade biology lesson from a Genetics unit

<table>
<thead>
<tr>
<th>Time</th>
<th>Author</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:12</td>
<td>Student1</td>
<td>yes both of the parents are homozygous but yellow cat is dominant and white cat is recessive</td>
</tr>
<tr>
<td>00:27</td>
<td>Tutor</td>
<td>Let me make sure I understand you - are you saying a white cat had to come from parents who both carry the recessive white fur gene?</td>
</tr>
<tr>
<td>00:36</td>
<td>Student2</td>
<td>because the orange color coat is more dominant than the white color coat</td>
</tr>
<tr>
<td>00:57</td>
<td>Student1</td>
<td>its true</td>
</tr>
</tbody>
</table>

4.3 Agree-Disagree Facilitation

We also present a conversational agent behavior based on the “Agree-Disagree” APT move. As the group discusses flows, the agent monitors the chat for student assertions that could be followed up by a check for agreement or understanding. This uses the detection method described above, but with a looser match threshold than that employed by the Revoicing behavior. This results in the detection of a greater number of candidate statements, and more opportunities for support than the Revoicing agent could afford. Not all of these detected statements will result in the agent triggering an Agree-Disagree move. Instead, when a candidate statement is identified, the Agree-Disagree component waits for the other students in the group to respond to it. If another student responds with an evaluation of their peer’s contribution (along the lines of ‘I agree’ or ‘I think you’re wrong’), but does not support their evaluation with an explanation, the agent will encourage this second student to provide one. If a student instead follows up with another candidate statement, the agent does nothing, leaving the floor open for productive student discussion to continue unimpeded, reducing the risk of over-scripting their collaboration. If the other students do not respond with either an evaluation or a contentful followup, the agent prompts them to comment on the candidate statement - for example, “What do you think about Student’s idea? Do you agree or disagree?” This interactive process is illustrated in Figure 4.1.

Table 4.3 shows an example of this support in the high-school biology setting. Times are given in seconds from the beginning of the excerpt. Rows marked in italics are the automatically detected labels the agent uses to motivate its facilitation moves. Student1 offers a contribution that is a candidate for evaluation. After 15 seconds of no students following up, the agent inserts an Agree-Disagree prompt. Student2 then responds very quickly, and then Student3 follows suit. Thus, the agent’s prompt seems to have successfully engaged the group in responding to Student1. The excerpt shown in Table 4.4 is drawn
from a unit of college Chemistry in intermolecular forces. This exchange is typical of a group interaction in this environment - Student1’s contribution at 17 seconds, although a candidate for evaluation, is not acted upon by the agent because Student3’s follow up pre-empts it. After 15 seconds following Student3’s statement without any sort of uptake by the group, the agent prompts the group to agree or disagree with it. Student2 offers a challenge, which leads to an extended back-and-forth between all three students.

Table 4.3: Agree-Disagree example in a 9th grade biology lesson from a Genetics unit.

<table>
<thead>
<tr>
<th>Time</th>
<th>Author</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00</td>
<td>Student1</td>
<td>the yellow cat is probably GG and the white is gg</td>
</tr>
<tr>
<td>00:15</td>
<td>Tutor</td>
<td>Do you agree with Student1? Why, or why not?</td>
</tr>
<tr>
<td>00:17</td>
<td>Student2</td>
<td>Gg,Gg,Gg,Gg.</td>
</tr>
<tr>
<td>00:47</td>
<td>Student1</td>
<td>I agree with Student2 for the possible outcomes</td>
</tr>
<tr>
<td>00:48</td>
<td>Student2</td>
<td>I agree with Student1</td>
</tr>
</tbody>
</table>

4.4 APT Feedback

A final behavior implemented using Bazaar and evaluated in this paper provides positive feedback for student APT. Here the goal is to offer encouragement when students themselves begin to act as APT facilitators with one another. Thus, rather than offer APT facilitation moves, the agent’s task here was to detect when students were doing it. Student input was matched against a list of hand-crafted patterns indicating APT moves including explanation, challenges, revoicing, and requests for others to provide each of the same. If a student statement matched, the agent publicly praised the student’s move, and (when appropriate) encouraged the other students to attend to and interact with that student. Rather than perform APT based facilitation itself, as the Revoicing behavior does, the Feedback behavior was meant to indirectly support the prevalence of APT in the discussions by encouraging students to take this facilitation role. An example is presented in Table 4.5 where moves detected as worthy of positive feedback are marked in italics, and the agent’s feedback is marked in bold.
Table 4.4: Agree-Disagree example in an undergraduate Chemistry unit on intermolecular forces.

<table>
<thead>
<tr>
<th>Time</th>
<th>Author</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00</td>
<td>Student1</td>
<td>ok lots of things to do...</td>
</tr>
<tr>
<td>00:13</td>
<td>Student1</td>
<td>first one</td>
</tr>
<tr>
<td>00:17</td>
<td>Student1</td>
<td>surface area is higher</td>
</tr>
<tr>
<td>00:20</td>
<td>Student2</td>
<td>arrow up?</td>
</tr>
<tr>
<td>00:22</td>
<td>Student3</td>
<td><em>ok, boiling pt will go up and vdW will go up for all of them consecutively... right?</em></td>
</tr>
<tr>
<td>00:37</td>
<td>Tutor</td>
<td>Do you concur with Student3? Why, or why not?</td>
</tr>
<tr>
<td>00:41</td>
<td>Student2</td>
<td>hmm not necessarily</td>
</tr>
<tr>
<td>00:47</td>
<td>Student1</td>
<td>area goes up for each</td>
</tr>
<tr>
<td>00:50</td>
<td>Student2</td>
<td>would it?</td>
</tr>
<tr>
<td>00:51</td>
<td>Student2</td>
<td>im not sure</td>
</tr>
<tr>
<td>00:56</td>
<td>Student3</td>
<td>yea for sure area goes up</td>
</tr>
</tbody>
</table>

Table 4.5: APT Feedback example in a high school Biology unit on diffusion

<table>
<thead>
<tr>
<th>Time</th>
<th>Author</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:41</td>
<td>Student1</td>
<td>it weighs more because there is more in it</td>
</tr>
<tr>
<td>08:45</td>
<td>Student2</td>
<td>starch is tested with a tube as the glucose is tested with a piece of paper like material</td>
</tr>
<tr>
<td>08:45</td>
<td>Tutor</td>
<td>Thanks for offering an explanation, Student1 :-)</td>
</tr>
<tr>
<td>08:46</td>
<td>Student3</td>
<td>the longer you leave the test strip in the water the darker green the strip gets and the more weight the glucose solution collects</td>
</tr>
<tr>
<td>09:22</td>
<td>Student1</td>
<td>Student3, wouldn’t it just show that there was more in it</td>
</tr>
</tbody>
</table>
Chapter 5

Automated Accountable Talk Facilitation in Action

The line of inquiry described in this chapter was prompted by the hypothesis that by incorporating intelligent agents to model, support, coach, and provide feedback for students using Academically Productive Talk (APT) moves, students will benefit in terms of learning and interaction. Note that we do not hypothesize that all APT moves are interchangeable. Rather, we manipulate the usage of different APT moves in order to understand better their separate and joint effects on measures of learning and interaction. The experiments presented in this chapter build on the early success of a form of APT, namely Revoicing support, in a study with 9th grade biology students [26]. The series of studies presented in this chapter (and also in a recent journal article [3]) served as a test of the generality of the effect across different learning contexts.

What we see in this series of studies is that the positive effect of APT facilitation behaviors is context specific. Generalized APT support would need insights into the contextual pre-conditions for the success of these facilitation strategies. The pattern of results across the studies begins to provide an empirical foundation for a more agile, more generalizable form of support that can use APT facilitation behaviors in a more nuanced, population sensitive way. Note that we are not claiming in our presentation of these studies that we already have this agile form of support. Rather, our investigations provide the initial empirical foundation for developing such an approach.

In addition to the learning gains analysis for each study, we present an automated process analysis technique that proves surprisingly accurate in identifying which interventions were most successful in each context. An automated measure that provides an indication of the relative success of alternative intervention strategies within contexts can be used to discover new associations between contexts and facilitation strategies in real time. Thus, we argue that beyond the insights into the individual contexts investigated in this series of studies, the results allow us to make cautious predictions beyond those contexts using the results from our process analysis.

These results are summarized in Table 5.2 at the end of this chapter.

5.1 Experimental Paradigm

In all four studies discussed in this chapter, which includes the foundational study [26] and three more recent [3], the instructional goal is for students to understand principles that explain causal mechanisms at a deep level. To that end, we prompt students for explanation in the context of group discussion with the goal that students will articulate and monitor their own reasoning, evaluate one another’s reasoning, and challenge one another. In all cases, students interact with their group members by logging into a chat room assigned to their group in the ConcertChat environment displayed in Figure 1 above, a discussion environment with a shared whiteboard [61].
5.2 Assessment

In all studies presented in this chapter, we employ both summative assessments in the form of pre/post domain-knowledge tests, as well as process assessments that measure the interventions’ success in eliciting more of the behaviors that mark effective collaborative learning processes. Thus the first analysis we do in all studies is to verify that learning took place between pre and post-test (using an ANOVA) and then to test for differences in learning between conditions (using an ANCOVA).

Beyond the learning gains analyses, we also do a process analysis. The specific interaction goal of APT interventions is to engage students in a more intensive exchange of explanations. More specifically, the desired contributions within these exchanges are what we referred to above as revocable assertions. By more intensive, we do not mean that students utter more explanations per se, but that the explanations they utter are directed towards building on those of their partner students. The motivation for attempting to achieve this was to raise the level of critical thinking and learning. Thus, in addition to a Pre/Post test measure of learning, a process analysis to verify that the intervention did its job is also important for evaluating our hypothesis. Anecdotally, we have observed that in some conversations, there were bursts of explanation behavior where this kind of intensive knowledge exchange was taking place. The purpose of our quantitative process analysis was to measure the extent to which this kind of bursty behavior was occurring within discussions as a result of the manipulation.

In order to accomplish this, the chat logs were segmented into intervals such that one observation is extracted per student for each interval. For young learners, we use 5 minutes as the interval since they type slower and take more time before responding whereas for older, more advanced learners, we use 2 minutes as the interval. In this way, we keep the average number of contributions per segment comparable between age groups. In each observation, we count the number of revocable assertions contributed by the student and the number of revocable assertions contributed by other group members. Conversations with more bursty behavior patterns should have a higher correlation between these two variables, which would signify that students are more active in the conversation when their partner students are also active.

Thus, for the process analysis, we evaluate the effect of condition on the correlation within time slices between occurrences of revocable assertions of a student with those of the other students in the same group. We used a multi-level model to analyse the results in order to account for non-independence between instances. We expect to see that the correlation is significantly higher in the condition with the intervention when the intervention is effective. We do the analysis separately for each independent factor within each study in order to contrast discourse behaviour between conditions. Specifically, we used what is referred to as a random intercept and slope model, which allows estimating a separate latent regression line for a student’s behavior in relation to that of their partner students within time slices. In this model, each student trajectory is characterized by a regression with latent slope and intercept. To do this analysis, we used the Generalized Linear Latent and Mixed Models (GLLAMM) add-on to STATA [62]. The dependent measure was number of revocable assertions by the student within the time slice. The independent variable was the number of revocable assertions contributed by the other students in the group within the same time slice. The condition variable was added as a fixed effect, and as an interaction term with the independent variable. A significant interaction between condition and independent variable in this case would indicate a significant difference in correlation between a student’s contribution of revocable assertions and that of their partner students. A positive difference would be indicative of an intensification of the interaction between students. A significant positive difference in intercept between conditions would indicate that the intervention raised the average number of revocable assertions within time slices.

5.3 Study 1: 9th Grade Diffusion Lab

In the initial published proof of concept regarding the effectiveness of APT agents at improving collaborative processes and learning [26], we ran a study in an urban high school, during a module introducing the concepts of selective permeability, diffusion, osmosis and equilibrium. The content was relatively new to the students, since they were at the beginning of a new unit in their course. Students worked together in a collaborative session for about 20 minutes, where task was to engage in a series of collaborative discussions
in which students make predictions, then make observations, and then explain why their predictions did or did not come to pass. Through this experience, the students would observe that glucose, water and iodine molecules all diffuse through dialysis tubing while starch molecules do not. The activity naturally lends itself to observing a variety of distinct cell models involving dialysis tubing containing an inside environment immersed in a beaker containing the outside environment. In each, a choice must be made for which liquid will be placed outside and which liquid will be placed inside. The collaborative task content, the macro-scripts that supported it, and the list of key concepts used for revoicing were all developed iteratively with feedback from teachers and content experts. An excerpt from the study that shows the agent offering an APT move in its first turn and a Feedback move in its second turn is displayed in Table 5.1. The tutor’s feedback move is triggered by Student1’s explanation attempt in the first turn. The tutor’s revoicing move is triggered by Student3’s contribution in the fourth turn.

The APT agent in this study provided both macro-scripting and micro-scripting support for student collaboration. The macro-scripting support provided a common task structure across conditions. While acting in the role of macro support provider, the APT agent provided instructions for the collaborative task, and introduced each step of the collaborative task, with the goal of controlling for time on task across conditions. This behavior is not displayed in the excerpt above. The micro-level support was meant to respond to the particulars of the conversation as it unfolded. Each experimental condition was defined based on which behaviors would trigger a supportive move, and what that move would be.

Table 5.1: Example of Feedback and Revoicing in the Dyke et al. study.

<table>
<thead>
<tr>
<th>Time</th>
<th>Author</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:41</td>
<td>Student1</td>
<td>it weighs more because there is more in it</td>
</tr>
<tr>
<td>08:45</td>
<td>Student2</td>
<td>starch is tested with a tube as the glucose is tested with a piece of paper like material</td>
</tr>
<tr>
<td>08:45</td>
<td>Tutor</td>
<td>Thanks for offering an explanation, Student1 :-)</td>
</tr>
<tr>
<td>08:46</td>
<td>Student3</td>
<td>the longer you leave the test strip in the water the darker green the strip gets and the more weight the glucose solution collects</td>
</tr>
<tr>
<td>09:22</td>
<td>Student1</td>
<td>Student3, wouldn’t it just show that there was more in it</td>
</tr>
<tr>
<td>09:26</td>
<td>Tutor</td>
<td>Would another way to say that be “indicators can prove that there was a change in concentration?”</td>
</tr>
<tr>
<td>09:31</td>
<td>Student2</td>
<td>wat Student3 said and starch can’t get any darker when purple and the water would be clear so no more</td>
</tr>
</tbody>
</table>

This study was run as a 2x2 between subjects factorial design in which the interactive support provided some behaviors in common across conditions, but other behaviors were manipulated experimentally. The first variable for manipulation was the presence or absence of the Revoicing behavior. The second variable was the presence or absence of the APT Feedback behavior, which is simply positive reinforcement when students were detected to engage in APT behaviour with one another. Students showed significant learning gains in all conditions, and there was a significant main effect of Revoicing such that students in the Revoicing condition learned significantly more between Pretest and Posttest, with an effect size of $\sigma$. There was no significant main effect of Feedback although there was a trend for it to have a negative effect. And there was no significant interaction between the two factors.

Despite the substantial literature supporting the effectiveness of APT in classroom discussions, it must be acknowledged that much is not known about the mechanism through which the complex intervention has done its work. This can only be determined through more fine-grained, careful experimentation. The treatment has always been complex involving multiple facilitation moves, used within whole classes, where a human teacher insightfully decides when and with whom to use each move. The series of controlled studies presented in this chapter was meant to begin to fill this empirical gap, in order to begin to build an empirical foundation for evidence-based design principles for development of effective APT-inspired dynamic support for collaborative learning in groups. The Dyke et al. study is the first study that demonstrated the effectiveness of Revoicing as support for collaborative learning with 9th graders, and
thus it forms the starting place for our series of studies investigating the generality of the effect in Chapter 5.

In order to compare the results from this study with those of the other studies, we present now the process analysis results from this study. The process analysis using the random intercept and slope model showed an interesting contrast between the Revoicing intervention and the APT Feedback intervention that is indicative of a possible explanation for the differential effect on learning during the collaborative activity. In the Revoicing condition (where there was a Revoicing agent to offer micro level support), we saw the pattern that we anticipated in conjunction with a positive learning effect in comparison with the Control condition (where there was no Revoicing agent). There was no significant difference in intercept between conditions, confirming that there was no difference in absolute number of revoicable assertions between conditions. More importantly, there was no significant correlation between the number of revoicable assertions of a student and that of his partner students in the Control condition where there was not a Revoicing agent. However, there was a significant interaction between the Revoicing condition variable and the number of revoicable assertions contributed by partner students ($R = .14, z = 2.03, p < .05$). This indicates that there was a significantly higher positive correlation between the number of revoicable assertions contributed by a student and that contributed by partner students in the Revoicing condition. Thus we do see evidence that in the Revoicing condition, the intervention had the effect of precipitating pockets of intensive discussion.

In contrast, with the APT Feedback intervention we see an entirely different pattern. In this case, there was a significant positive effect on the intercept associated with the APT Feedback condition, indicating that students contributed significantly more revoicable assertions in the APT Feedback condition. However, there was a marginal interaction between condition and the number of revoicable assertions, this time with a negative coefficient ($R = -.16, z = -1.87, p = .07$). This indicates that while students were talking more, they were interacting with one another less intensively, which is consistent with the finding of no effect on learning. A possible explanation is that the Feedback agent elicited interaction between students and itself while the Revoicing agent elicited interaction between students, which was the goal.

### 5.4 Study 2: 9th Grade Genetics

![Punnett Square problems from 9th grade genetics unit](image)

The second study was conducted within the same course where the first study was conducted, but two months later, in a unit on Genetics. The study was carried out during a module specifically introducing the concept heredity, and the use of Punnet squares as a tool to reason about the inheritance of single traits. At the time of the study, the material was somewhat familiar to the students since they were towards the
end of the unit by the time the study took place. In the collaborative activity that lasted for about 20 minutes, student groups were presented with a set of three problems and asked to reason about the physical and genetic traits of the hypothetical parents of a set of sibling organisms. Specifically, in each problem, students were shown a litter of eight kittens that varied in fur color (either orange or white), and were instructed to identify the genotypes and phenotypes of the parents, and to explain their reasoning to their teammates. This sort of “backwards” reasoning had not been explicitly addressed in the course to date - students only had prior experience with “forward” reasoning from given parental traits. The mystery parents were presented as the inputs to an unpopulated Punnet square, as shown in Figure 5.1. As an incentive, students were told that the best team, determined by a combination of discussion quality and post-test scores, would be awarded with a modest prize of food. Each of the three tasks was progressively harder than the last in that fewer clues about the parent’s identities were included. The collaborative task content, the macro-scripts that supported it, and the list of key concepts used for revoicing were all developed iteratively with feedback from teachers and content experts.

5.4.1 Participants

This study was conducted in the same seven 9th grade biology classes of an urban school district that the first study was run in, only two months later. The classes were distributed across two teachers (with respectively 3 and 4 classes) for a total of 78 consenting students, who were randomly assigned to groups of 3. Groups were randomly assigned to conditions.

5.4.2 Experimental Manipulation

In this study, only Revoicing behaviors were manipulated experimentally. The APT Feedback that was evaluated in the first study was not repeated in the second study since it did not lead to a positive effect with this student population in that study. In both conditions of this study, the agent provided the same macro level support by guiding the students through the activity using the same phases introduced in such a way as to control for time on task. Only the micro-level support varied between conditions.

5.4.3 Study Procedure

Just like in the first study, the students first participated in a normal class lesson on genetics as part of the course curriculum. At the end of the period, they took a pre-test. The pre-test included four multiple-choice questions testing the students’ ability to use Punnet squares to reason about the likelihood of genetic and physical traits of children based upon the traits of the parents, and one open-ended question designed to elicit explanation of reasoning about parental identity based upon the physical traits of offspring.

In the next class period, the students participated in a 20 minute collaborative computer-mediated activity during which the experimental manipulation took place. The students did the activity in groups of three students, scaffolded by conversational agents. Students within classes were randomly assigned to groups and then groups to conditions. As in the first study, this activity was introduced by a cartoon depicting the use of APT and a reminder of the basic science principles underlying the activity, in this case principles of simple inheritance. At the end of this second phase, the students took a post-test of the same design as the pre-test, although with different characteristics and genotypes presented in each problem. Results

As in the Diffusion Lab study, we evaluated pre-to-post test learning and the effect of condition on learning and on the collaborative process. However, the material appears to have been too easy for the students. Post-test scores were higher on average than pre-tests scores, but not significantly. And although the trend was for students in the Revoicing condition to learn more than students in the Control condition, the difference was not significant or even marginal. Thus we do not elaborate on the learning gains analysis here.

While the learning gains analysis does not allow us to draw new insights about learning, we can observe how the collaborative processes play out with the same student population used in Study 1, but with material that appears to be less challenging for them. The process analysis using the random intercept
and slope model showed an interesting contrast between this study and the Diffusion lab study. Similar to the Diffusion study, there was no significant difference in intercept between conditions, confirming again that there was no difference in absolute number of revocable assertions between conditions. This time, however, there was a significant correlation between the number of revocable assertions of a student and that of his partner students in both conditions ($R = .31, z = 3.59, p < .001$), and no difference in slope between conditions. Thus, we have confirming evidence that there was no difference in effect between conditions. Students were interacting productively in both conditions regardless of support, possibly because the material was easy for them and thus they may not have needed the revocing support.

5.5 Study 3: Freshman Engineering Design

As a second replication of the successful Diffusion Lab study, we ran a study in a Freshman Engineering Design course at a selective private university. The material presented in the study was relatively familiar to the students. The experimental manipulation was identical to that of Study 1, including both the APT Feedback manipulation and the Revocing manipulation.

5.5.1 Participants

109 mechanical engineering students participated in the experiment, which was held over six sessions spread evenly between two days. Students were grouped into teams of three or four individuals. The number of three person and four person groups was roughly evenly distributed between conditions. In each session, the groups were evenly distributed between the three conditions. The two days of the experiment were separated by two weeks.

5.5.2 Experimental Procedure

Each session started with a follow-along tutorial of computer-aided analysis where the students analysed a wrench they had designed in a previous lab. A pre-test with 11 questions (7 multiple choice questions and 4 brief explanation questions) was administered after the analysis tutorial. The experimental manipulation happened during the Collaborative Design Competition after the pre-test. Students were asked to work as a team over 90 minutes to design a better wrench taking three aspects into consideration: ease of use, material cost and safety. Students were instructed to make three new designs and calculate success measures for each of the three aspects under consideration. As part of this process, students occasionally were requested to make predictions and explain them, however, it should be noted that this task was somewhat less conceptually oriented than that used in the other studies.

5.5.3 Results

The results of this study were strikingly different from the two conducted in 9th grade Biology. In particular, rather than achieving a positive effect, the Revocing manipulation had a significant negative effect on learning within the APT Feedback condition with this more advanced population of learners.

As in the earlier studies, we began our analysis by first verifying that students learned between pre and posttest. For this analysis, we treated Test as a repeated measure, with Pre and Post being the two time points. We conducted an ANOVA test with Test as the dependent variable. Time point, Revocing, and Feedback were independent variables. We included all two-way interaction terms as well as the three-way interaction term. There was a significant main effect of Time point $F(1,210) = 9.28, p < .005$, demonstrating that students learned. None of the interaction terms were significant. Thus students learned between pre and posttest regardless of condition.

Next we tested for differences in learning between conditions. For this analysis, we conducted an ANCOVA with Post-test as the dependent variable and Pre-test as a covariate. Revocing and APT Feedback were the two independent variables. We also included the interaction term in the model. Here there was almost no effect of APT Feedback $F(1,104) = .03, p = .87$. There was a trend for a negative effect of the Revocing manipulation $F(1,104) = 2.22, p = .13$. The interaction between APT Feedback
and Revoicing was not significant, however, it should be noted that within the APT Feedback condition, there was a significant negative effect of Revoicing ($p < .05$). Thus, there is some qualified evidence of a potential detrimental effect of Revoicing with this population.

Consistent with the negative trend, the process analysis using the random intercept and slope model showed an interesting contrast with the earlier studies when we evaluated the effect of the Revoicing manipulation. Similar to the earlier studies, there was no significant difference in intercept between conditions, confirming again that there was no difference in absolute number of revocable assertions between conditions. There was, however, a significant correlation between the number of revocable assertions of a student and that of his partner students in the control condition ($R = .1, z = 3.7, p < .001$), as well as an interaction between condition and slope. In contrast to the Diffusion study where we saw a positive effect of revoicing both on learning and on the slope, here we see a negative impact on slope based on the correlation on the interaction term. This echoes the trend for a negative effect on learning ($R = - .1, z = 2.4, p < .05$). Thus, we have confirming evidence that there was a negative impact of the Revoicing manipulation with this population. When we do the same analysis to evaluate the effect of the APT Feedback condition, we see no effect of any variable.

5.6 Study 4: Freshman Honors Chemistry

In the final study, published as a conference paper [1], we tested the hypothesis that one reason why Study 3 was not successful was that the students did not need support in making themselves clear. Instead, we hypothesized that instead of support for basic articulation of ideas, they needed support to the next step of challenging each other’s reasoning. We consider this study to be a good comparison case to Study 3 because the student population was similarly university level from the same selective private university, and the material was similarly relatively familiar to the students.

The collaborative task, which lasted for about 90 minutes, focused on intermolecular forces and their influence on the boiling points of liquids. For each problem in the activity, students were asked to predict whether a given substance would have a higher or lower boiling point than two of its relatives, explaining their reasoning about the set of molecules in terms of their structure and the forces at play. Each problem of this sort was followed up by revealing the actual boiling point of the mystery molecule, and asking students to revisit their predictions and explanations in light of the new data. A liquid’s boiling point can be influenced simultaneously by a number of different intermolecular forces, each of which arises as a consequence of the molecules’ particular structural attributes. Correctly identifying the pertinent structural features of molecules and reasoning about how they will affect the liquid’s boiling point is a non-trivial and multi-faceted task. Because multiple types of intermolecular forces influence liquids’ boiling points, we employed the Jigsaw technique [6], assigning students within each group to read individually about one of three forces that contribute to a molecule’s boiling point. This division also provided intrinsic motivation for collaboration, as the task could not be completed without knowledge from each of the student experts.

5.6.1 Participants

The participants in our study were first-year undergraduate students studying intermolecular forces in an Honors Chemistry course. Students were randomly assigned to groups of three or four, and then groups were randomly assigned to conditions. The balance of three and four person groups was even between conditions, and there was no effect of team size on any of our dependent measures. All students in the course were required to participate in the online exercise for course credit, but they had the option of not consenting for their data to be included in our research. Thus, we only report results for consenting students. Altogether, our analysis includes data from 18 students from 6 different groups, which is 9 students and 3 groups in each condition.
5.6.2 Experimental Manipulation

Our experimental design was a simple 2-condition between-subjects design where teams were assigned randomly either to the Agree-Disagree condition or the Control condition. Both conditions were identical except for inclusion of the Agree-Disagree facilitation move by the agent. Thus, both conditions benefitted both from macro-level and micro-level script based support. In the Agree-Disagree condition, whenever the agent was not engaged in a directed dialog, it was receptive to opportunities to dynamically offer support using the Agree-Disagree behavior, discussed above.

5.6.3 Experimental Procedure

The experimental procedure was simple. Students took a pretest, then participated in pairs in the online collaborative activity, and finally completed a post-test. Pre and post tests were used to measure learning during the collaborative exercise.

5.6.4 Results

Our hypothesis was that the introduction of the Agree/Disagree agent would intensify the interaction between students, which might increase critical thinking, and subsequently increase learning. Our analysis offers qualified support for the hypothesis.

As before, we began our analysis by first verifying that students learned between pre and posttest. For this analysis, we treated Test as a repeated measure, with Pre and Post being the two time points. We conducted an ANOVA test with Test as the dependent variable. Time point and Revoicing were independent variables. We included the interaction between Time point and Condition as well. There was a significant main effect of Time point $F(1,31) = 7.58, p < .01$, demonstrating that students learned. The interaction term was not significant. Thus students learned between pre and posttest regardless of condition. As before, to evaluate the effect of condition on learning, we used an ANCOVA with posttest as the dependent variable, pretest as a covariate, Condition as an independent variable. In this analysis, there was a marginal effect of Condition on learning ($F(1,11) = 1.82, p < .1$, effect size .55$\sigma$), such that students in the Agree/Disagree condition learned more. The effect was moderate.

Next we examined the intensifying effect of the intervention on the interaction between students using the same random intercept and slope model approach used in the earlier studies. The analysis showed the pattern that we expected. There was no significant difference in intercept between conditions, confirming that there was no difference in absolute number of revoicable assertions between conditions. More importantly, there was no significant correlation between the number of revoicable assertions of a student and that of his partner students in the control condition where there was not an Agree/Disagree agent. There was, however, a significant interaction between the condition variable and the number of revoicable assertions contributed by partner students ($R = .14, z = 2.03, p < .05$). This suggests that there was a significant positive correlation between the number of revoicable assertions contributed by a student and that contributed by partner students in the Agree/Disagree condition. Thus we do see evidence that the intervention had the effect of precipitating pockets of intensive discussion.

5.7 Discussion

The pattern of results across studies is consistent with what we expected to see given the connection between types of transactive discussion behavior and how they are related to the three different discussion facilitation behaviors we explored in the current body of work. In particular, we contrasted Revoicing, which is meant to elicit self-oriented, consensus-oriented transacts, which we have argued should be less demanding and to some extent logically prior to other-oriented, conflict-oriented transacts, which are elicited by Agree-Disagree facilitation moves. It would therefore be consistent to expect that Revoicing moves would be most needed by younger, less sophisticated learners, whereas Agree-Disagree moves would be more appropriate for more advanced learners. In prior studies of the effect of transactivity on learning[7], the effect was only observed in material that was difficult for learners, thus we would expect that learners
who were close to mastery would not benefit substantially from APT. Thus, where material is easy for learners, we would not predict a difference between conditions where we test APT in comparison with other facilitation behaviors or even no facilitation.

In Study 1, where we test Revoicing against Feedback for APT with young learners on material that was difficult for them, we observe a positive effect of Revoicing. In Study 2, we test Revoicing again, but this time with material that was easy for the students. Here there was no significant difference between conditions. This contrast is consistent with what we argued above. It is true that since the group of learners was the same in the two studies, the difference in effect could have potentially been related to the fact that the students were already familiar with the support agents. It is clear, however, that re-exposure to the same manipulation does not completely explain the difference in results across these two studies. In the first study, we observed a significant pre to post test gain across all conditions, including the condition where no support was offered beyond the macro level structuring of the activity. In the second study, no significant pre to post test gain was observed in any condition. Rather, both pre and post test scores were high across conditions, which highlights the fact that the material was easy for the students.

Studies 3 and 4 involve more advanced learners on material that was moderately familiar to them. More advanced learners are already good at articulating their own ideas. Thus, Revoicing support is unneeded support for them. Rather, they need to be pushed beyond that to connect to the reasoning of their partner students. We expect then not to see a positive effect in Study 3 where we test Revoicing on these advanced learners, and we do expect to see a positive result with Agree-Disagree, which we test in the final study. And we do see this.

The pattern of results with learning gains is as expected from prior work. What is more striking is the picture that emerges when we compare the pattern of results from the learning gains analysis with that from the process analysis. What we see from the series of studies presented in this chapter is that the effect of condition on learning gains and on collaborative process provide largely converging evidence across studies. This convergence highlights the value of the simple form of process analysis presented in this chapter for evaluating in process effect of collaboration support. It shows that this process analysis can be used to gauge whether an intervention is working appropriately with a group of learners. If the process analysis indicates that the strategy is not a good match for the learners, the strategy can be adjusted. The new strategy can then be evaluated in process the same way, and further adjustments can be made. Thus, this simple automated process analysis technique could form the foundation for a new, more agile approach to dynamic support for group learning where the strategy itself can adapt to the needs of the population of learners.
5.8 Conclusions

The work here described has laid an empirical foundation for a research agenda for a new generation of dynamic and agile support interventions to improve collaborative learning. As we have demonstrated through an integration of results from four experimental studies, the effects of dynamic support vary based on the ability level of learners as well as the nature of the material itself. Human instructors are highly agile in their use of facilitation techniques (including Academically Productive Talk) along many dimensions, including group assignment, selection of facilitation moves, timing, and sequencing. We therefore argue that achieving a higher level of agility is needed to move to the next stage - agility in terms of selection of students to target, selection of interaction strategies, and timing. While the results presented in this chapter are compelling, it would be more compelling to examine the contrast between multiple different strategies within the same study. This proposed work is described in Chapter 6.

Agility comes with challenges from an experimental standpoint, however. As mentioned in the architecture discussion in Chapter 3, Bazaar’s flexible approach to interactive script integration allows a variety of scripting paradigms to be implemented, with varying effects on the agents’ internal and external validity. For example, specifying high priority and rigid constraints on macro-scripted actions, alongside low priority for dynamic feedback, produces an agent configuration with high internal experimental validity. In such a configuration, macro-script stages reliably occur at specified intervals, guaranteeing that each group of students interacting with instances of the agent engage with each stage of the script (and the associated learning opportunities) for the same amount of time. However, this comes with a loss of agility, and the potential for lost opportunities for natural collaborative conversation. The beginning of a new script phase may cut off an ongoing student conversation, or deny another component’s chance to complete a follow-up move. On the other hand, if the dynamic components are configured to reserve more follow-up time after their behaviors are enacted (or the macro-script is configured to wait for a period of inactivity before proceeding), there’s greater opportunity for natural flow and resolution in student and agent interactions. This lends a greater external validity to the experience, but with greater variability in timing and experience between instances.

The technical approach presented in this chapter enables a wide variety of strategies to be implemented. The work completed to date provides the seeds of the needed empirical foundation. We offer this set of results as an argument in favor of a larger, more thorough and systematic investigation of the space of possibilities, as described in Chapter 6.
Chapter 6

Proposed Work

The work completed to date constitutes an initial exploration of agent-based discussion facilitation for collaborative learning. However, the results are qualified by limitations of the studies, and by their piecemeal nature. We have only evaluated the effects of individual facilitation moves in fixed collaborative settings, without fully investigating their interactions or their situational suitability. How does a student’s transactive competence or in-domain skill-level influence the effect of facilitation strategies from each of the “levels” suggested by the frameworks of transactivity and Academically Productive Talk? How do the efficacy of facilitation strategies vary over the course of an evolving conversation? Can the agile selection of such strategies be successfully automated? We propose a series of further studies and analyses to address these remaining questions.

A timeline for this proposed work is presented in Section 7.2.

Hypotheses

1. The most suitable facilitation strategies for a given setting will vary with such contextual factors as age, content knowledge, task difficulty, and self-efficacy. In particular, we hypothesize that they will vary in a manner consistent with the transactivity and APT frameworks - lower-level strategies will be more suitable for less experienced students struggling with the content, and higher-level strategies will be better for more experienced groups that are comfortable with the content.

2. During a given collaborative session, the most suitable facilitation strategies will vary with the conversational trajectory of the group. We anticipate that this variation will be consistent with our theoretical framework - groups that show evidence of building on and transforming each others’ ideas will benefit from higher-level strategies (or less intervention overall), while groups that have trouble listening to each other will do better with strategies supporting more basic transactivity.

3. Automated, agile facilitation can effectively support student learning and collaborative processes. In particular, we expect that such a system will be more effective than one using a static selection of facilitation strategies.

6.1 Proposed Study: Full-Spectrum Facilitation

As framed in Chapter 5, there is need to examine which facilitation strategies are most effective in a given situation - both at a macro-level, depending on a group’s perceived skill level and and background, and at a micro-level, as students and groups demonstrate capacity for self-regulation and transactive talk.

In order to validate the observations of the previous studies, and to test Hypothesis 1, we will return to the college chemistry task. Facilitative agents using one or more strategy from across the levels described in Chapter 2 will support the students as they complete the same collaborative chemistry assignment described in Section 5.6. As in those completed studies, students will be given pre-tests and post-tests to assess their content knowledge, motivation, and self-efficacy, and process analyses will reveal their
collaborative behavior. In addition to verifying the effects of these facilitation strategies, we will be able to analyze how these different moves affect student performance, relative to their entry level.

Data from a similar study (with only two kinds of facilitation strategies from opposite ends of the transactivity spectrum) was collected from the general chemistry course at a top-tier private university, and is still under analysis. This replication and extension will be conducted with the participation of chemistry faculty at the same private university, and also at a large state university. In addition to student and group-level comparisons, we will be able to compare the effectiveness of the Agree-Disagree and Revoicing moves in this setting against the honors-level private university context used for the first chemistry study.

We plan to evaluate facilitation strategies at each of the four levels of the Accountable Talk framework (which also correspond to increasing levels of transactive reasoning) separately. In order, these are Revoice, Ask For Restatement, Press for Reasoning, and Agree/Disagree.

6.2 Proposed Study: What Would Teachers Do?

Our work to date has employed well-motivated mechanisms for triggering facilitation support, but these are ultimately manually defined rules crafted by a single hand, relying on a small set of convenient signals. We propose to compare the behavior of our system against the expectations of seasoned high school science teachers and facilitators with experience in Accountable Talk facilitation. These annotators will label missed opportunities for facilitation in both supported and unsupported chat transcripts, and rate the suitability of actual facilitation moves delivered by agents and classroom teachers. Such annotations will provide a validation of our present approach, and can serve as gold standards for models of facilitation based on expert judgement.

In addition to the experts’ labeling of opportune moves, we also intend to annotate our data with several theoretically motivated and automatable signals that may be used to anticipate or evaluate the suitability of facilitation moves in future facilitator agents, as described in Section 6.3. Table 6.1 describes several possible feature sets.

<table>
<thead>
<tr>
<th>Label</th>
<th>Reference</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transactive Moves</td>
<td>Berkowitz and Gibbs [13]</td>
<td>Can student transacts (or the lack thereof) anticipate the need for certain kinds of facilitation?</td>
</tr>
<tr>
<td>Target Statement</td>
<td>Adamson et al. [11], Fernando and Stevenson [29]</td>
<td>This is the primary trigger used by the current APT Agents</td>
</tr>
<tr>
<td>Similarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student and Group Participation</td>
<td>Adamson et al. [11], Kumar [47]</td>
<td>This encompasses the secondary triggers used by the current APT Agents</td>
</tr>
<tr>
<td>Negotiation Moves</td>
<td>Martin and Rose [52], Mayfield et al. [54]</td>
<td>For the purpose of tracking other signals useful to predict facilitation opportunities, like confusion/consensus</td>
</tr>
<tr>
<td>Active Learning Moves</td>
<td>Bharadwaj &amp; Singh (ITS 2014, submitted)</td>
<td>Built on top of negotiation, promising initial results submitted to ITS 2014</td>
</tr>
</tbody>
</table>
6.3 Proposed Work: Modeling Conversational State to Predict Facilitation Opportunities

Automatically extracted text features and potentially automatable discourse features can be used to model conversations and make predictions about their outcomes \[39, 53\]. In addition to lexical and syntactic features, we will explore modeling conversational state based on sequences of discourse act labels. In pursuit of Hypothesis 2, we expect such features and models can be used as inputs to new models that can identify appropriate opportunities for strategic facilitation.

6.3.1 Active Learning Annotation

To represent features above the contributions of individual lines of dialogue, we can refer to established frameworks for conversational analysis. In Barros et al.’s work, a set of attributes for qualitative conversational analysis is proposed \[11\] based on a set of six sentence-opening moves. This is similar to the scheme used by Soller \[75\]. We combine Barros’ two types of proposal and consider just five types of “Active Learning” moves:

- Proposals (PR) begin a sequence and introduce a new concept or idea.
- Questions (QU) target proposals and question them.
- Clarifications (CL) are elaborations on proposals, or answers to questions.
- Agreements (AG) show agreement or assent between speakers in a sequence.
- Remaining contributions are Comments (CM); including topic statements, floor grabbing moves, pauses, etc.

In other work using discourse act labels to predict student behavior, the assignment of turn labels has relied on student inputs being constrained to a fixed set of sentence-openers. In our approach, the students are not thus fettered, and we instead rely on annotation of free text. To allow this flexibility, we adapted a coding manual based on the systemic functional linguistics “Negotiation” framework \[52\], describing the flow of information and action within a conversation. Recent work has shown that Negotiation annotation can be automated for freeform chatroom conversations \[54\]. With an eye toward automation, we adapt Mayfield’s coding manual, converting Negotiation labels to Active Learning moves using heuristics.

From these annotations, we can represent sequences of labeled turns as inputs to machine learning algorithms. As a starting point, we will use sequences of three consecutive labels, extracted from the sequence of labeled turns, as a feature for our group and student tasks. In the case of per-student outcome prediction, each tag is differentiated based on who (relative to the student in question) is speaking - either the student herself, or another participant. For example, PRs is a proposal issued by this student, CLo is a clarification by another student, and so on. We can consider this representation both on its own and as a supplement to textual features.

6.3.2 Contrastive Hidden Markov Models

As a more sophisticated differentiator of conversational structure, we will use Hidden Markov Models \[63\] to model variation between successful and unsuccessful students. In this context, the hidden states may correspond to a student’s intention when contributing a new turn to the dialogue. By analyzing sequences of observed labels, HMMs can discover these unobserved states statistically. In our work, we may use Negotiation or Active Learning labels as observed states. Alternatively, we may employ block HMMs to work directly with distributions of text features within each turn.

Following Soller et al. \[76\], we can train two HMMs on sequences drawn from subsets of a collaborative chat corpus - for example, one using the sequences from a set of students with the highest residuals and the other using a set of students with the lowest residuals. We may also use other criteria, the favored frequencies accountable talk moves as indicated by human annotators in Section 6.2 to divide the population.
A comparison of estimated likelihoods from each model, as applied to an unseen conversation, can be used to classify that new conversation as belonging to the more likely model’s training population. This may allow an automated system to differentiate high- and low-performing students, or groups likely to benefit from a certain set of APT moves. We make no particular presumptions about the meanings of specific hidden states [25], although we expect to see meaningful patterns relevant to the collaborative discourse context.

By building a model based only on the early stages of a collaborative activity, we can anticipate a group’s trajectory during an ongoing activity. For example, groups (or individual students within groups) that are forecast to be lower-performing may benefit from increased facilitation at a lower level of transactivity. If their classification shifts with the addition of continuing conversational data, the selected facilitation strategies can shift accordingly.

6.3.3 Silent Signals for Facilitation

In addition to textual features and sequences of discourse labels available from group chat, a learning environment may be instrumented to allow student actions (for example, interactions in a shared geometry sandbox [79]) to be recorded. Analysis of student actions in a virtual chemistry laboratory has revealed variations in problem-solving strategies that reflect affect, and influence task success [8]. If such data is available, environmental actions may also be a source of information about a student or group’s collaborative state. The actions may also be used to trigger facilitation strategies. Asking a student to explain her reasoning for an enacted (but unspoken) problem-solving step, or offering a chance to challenge a risky or ill-motivated procedure, provides an opportunity for externalization and reflection very much in line with the principles of APT.

As the task and environment employed for the current set of studies does not offer such instrumented interaction, an investigation in this direction would require the integration of a Bazaar agent into a new learning activity. We have the opportunity to collaborate with researchers on a new study involving the aforementioned virtual chemistry laboratory [28], and may have other opportunities through recent partnerships, but the details remain unspecified. Bazaar’s modular design and simple configurability makes deploying a conversational agent into an extant collaborative task a fairly straightforward endeavor, developing a reliable collaborative task from the ground up is a much larger undertaking, somewhat out of the scope of this proposal. Depending on the availability and adaptability of action-instrumented collaborative learning tasks, explorations along this line may remain preliminary.

6.3.4 Evaluation

Following Hypothesis [2] we anticipate that modeling conversational state can effectively predict opportune moments for certain facilitation strategies. The performance of such models can be validated by comparison to the expert annotations. In cases where the data already includes proposed or enacted moves by a facilitator agent, or where we can simulate the existing agents’ behavior on unlabeled data, we can also compare this new model’s agreement with these ratings with that of the present system of hand-crafted rules. Further, we expect that the features and patterns learned by such a model will be interpretable within and consistent with the framework described in Chapter [2].

6.4 Proposed Study: Tying it Up With a Bow

As a capstone to this thesis work, we propose to deploy a facilitator agent powered by this new model in one of the same settings we’ve studied in the past (quite likely college chemistry), and compare the best-performing configuration of the old system against this new model (and both against no facilitative support) in terms of learning gains and collaborative process measures. We hypothesize that the agile facilitation system will more comprehensively support students in collaborative discussion, and this will be evident both in content learning and via the process measures described in Chapter [5].
6.5 Power Tools and Generalization

In order to test the generality of this approach, we would like to explore its application to domains that dwell farther from collaborative activities in high-school biology and college chemistry and engineering. Recent work exploring dialogue-powered personal assistants has addressed task planning as well as various social aspects of dialogue [51, 90]. Social supports and support for unexperienced users has been integrated into such assistants [33], and sequential models based on participant actions have been used to predict emotional state and to power backchanneling in embodied agents [59]. However, less attention has been paid to how these models might vary the nature of dialogic supports, in response to the users’ demonstrated needs and experience.

The specific Academically Productive Talk moves that are effective in group learning contexts may not be directly applicable in other domains. However, the general principles and technologies are likely to be transferable. We propose an initial exploration toward agile conversational support for task assistance, in either individual or group contexts. We imagine supporting users completing some multipart task that may draw upon multiple skill proficiencies. An effective application of agile conversational support in this context would adapt the degree and specificity of instruction, according to models of user confidence and confidence in each skill, with reference to the urgency or criticalness of each part of the task. Note that this thesis will not focus on task planning in any depth, but the conversational support we do focus on may fit nicely within such a system.

Working with collaborators and advisors at the Bosch Pittsburgh research office, we propose to integrate conversational support into a prototype system for dialogic task assistance, with a task or set of tasks that may involve home repair or kitchen cookery. Using data collected from an initial piloting phase without conversational support, we’ll investigate our framework’s applicability to modeling user state in this domain. Will sequential models of discourse between an individual and the system be able to predict outcomes in the same way that models of group discourse can? How do APT principles translate to moves suitable for adult users? Are challenges and opportunities for reflection productive to task success or skill-learning? While these questions cannot be fully addressed in the scope of this proposed exploration, we hope to establish a framework for future research in this area.
Chapter 7

Summary and Timeline

7.1 Summary of Contributions

**Advanced Architecture for Dynamic Conversational Agents**

Bazaar, as an open-source architecture for authoring dynamic, multi-policy conversational support agents, is a significant contribution to the fields of computer-supported collaborative learning and dialogue systems. The architecture’s high-level configurability, modularity, and advanced behavior orchestration methods allow for the rapid composition and deployment of conversational agents in collaborative environments, as research platforms and teaching tools. (Chapter 3)

**Framework for Agile Facilitation**

We provide an empirical foundation for a context-responsive approach to supporting group learning, rooted in the literature of transactivity and classroom facilitation. (Chapters 2, 5, 6)

**Practical Models for Automatic Facilitation**

Our completed work establishes the benefits of automated facilitation of small-group collaborative learning tasks using real-time conversational analysis, and underscores the need to take contextual factors into account when selecting facilitation strategies.

Our proposed work will extend this contribution by examining the potential of facilitation systems that dynamically account for such factors, considering both the macro-level context of group and task composition, and by automatically monitoring the micro-level evolution of conversational state within a collaborative group. (Chapters 2, 4, 9)

7.2 Timeline

While the timeline presented in Table 7.1 is by necessity guesswork, it provides a reasonable set of milestones for the work required to complete, document, and defend the proposed work.
<table>
<thead>
<tr>
<th>Completion Date</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb/March 2014</td>
<td>(scheduled) Run full-spectrum facilitation study with general chemistry students at private university (Section 6.1)</td>
</tr>
<tr>
<td>April 2014</td>
<td>(scheduled) Run full-spectrum facilitation study with general chemistry students at state university (Section 6.1)</td>
</tr>
<tr>
<td>Spring, Summer 2014</td>
<td>Label data with positive and negative examples of opportunities for APT facilitation, as marked by teachers and other facilitation experts (Section 6.2)</td>
</tr>
<tr>
<td>Summer, Fall 2014</td>
<td>Modeling student and group conversational trajectories (Section 6.3)</td>
</tr>
<tr>
<td>Summer 2014</td>
<td>Exploring Generalization (Section 6.3)</td>
</tr>
<tr>
<td>Fall 2014</td>
<td>Compare new agile facilitation model with existing hand-crafted rules in final study (Section 6.4)</td>
</tr>
<tr>
<td>Fall 2014 to Winter 2015</td>
<td>Write. (Chapters 1-7)</td>
</tr>
<tr>
<td>Spring 2015</td>
<td>Defend! (Chapters 1-7)</td>
</tr>
</tbody>
</table>
Bibliography


[26] Dyke, G., Adamson, D., Howley, I., Rosé, C.P.: Enhancing scientific reasoning and explanation skills with conversational agents. IEEE Transactions on Learning Technologies (2013) [1] [2.1] [3.1] [3.2] [3.3] [5.1] [5.3]


chat. SIGDIAL 2012 (2012) 6.1 6.3.1


[64] Raven, J.: Raven’s progressive matrices (rpm). Los Angeles, CA (1992) 2.3


