Gaining Insights from Sociolinguistic Style Analysis for Redesign of Conversational Agent Based Support for Collaborative Learning

Iris K. Howley, Rohit Kumar, Elijah Mayfield, Gregory Dyke & Carolyn P. Rosé

Abstract: Data from an early stage of development of conversational agent based support for collaborative learning provides an ideal resource for demonstrating the value of sociolinguistic style analysis paired with time series visualizations as part of an iterative design process. The methodology illustrated in this chapter was introduced in earlier publications focusing separately on the sociolinguistic style analysis (Howley & Rosé, 2011; Howley, Mayfield, and Rosé, 2013) and the time series visualization using the Tatiana tool (Dyke, Kumar, Ai, Rosé, 2012). However this chapter is unique in its application to data that is at such an early stage in a development process. The data is admittedly raw, and contains many examples of interaction gone awry. Nevertheless, the value in this analysis is in a demonstration of what insights can be gained through detailed stylistic analysis of conversational behavior that informs the next steps of intervention development.

Gaining Insights from Sociolinguistic Style Analysis for Redesign of Conversational Agent Based Support for Collaborative Learning

Iris K. Howley, Rohit Kumar, Elijah Mayfield, Gregory Dyke & Carolyn P. Rosé

Introduction

This chapter presents an analysis of collaborative learning interactions that uses the same approach to that employed in one of the analysis chapters of in the PLTL data section of this volume (Howley, Mayfield, Rosé, & Strijbos, this volume). As such, the relationship of this analysis and the dimensions of multivocality are similar as described in that chapter, although here we take a distinctly troubleshooting orientation. In particular, we similarly assume that collaborative learning processes are an integration of three orthogonal dimensions, namely cognitive, relational, and motivational. However, we adopt the cognitively focused dimension as our success metric, and use the other two more socially oriented dimensions to examine how collaborative processes that support the cognitive dimension went awry and explain a lack of success. Similar to the earlier analysis in the PLTL Chemistry section, we assume that each dimension can be operationalized as a set of mutually exclusive codes, each of which are defined at the level of an individual contribution to a conversation. Overall, the purpose of the analysis here is more specifically to reveal insights leading to suggestions for redesign of an intervention. Thus, in this analysis, pivotal moments are those key events in the interaction that lead to negative impact on collaborative processes and therefore should be avoided in a redesign of the intervention.

While theories of computer supported collaborative learning are many, there is a considerable consensus about what types of group interactions are desirable (Berkowitz & Gibbs, 1983; Teasley, 1997; Weinberger & Fischer, 2006; Suthers, 2006). It is widely acknowledged that groups do not operate at an ideal level without support. Thus, researchers in the area of scripted collaboration have worked to develop design principles to guide development of support that elicits the kind of group behaviors that are valued within the CSCL community (Dillenbourg, 2002; Kollar, Fischer, & Hesse, 2006; Kobbe, Weinberger, Dillenbourg, Harrer, Hämäläinen, Häkkinen, & Fischer, 2007). Through design-based research studies and experimental studies, theories of scripted collaboration have developed, which grapple with issues such as how much constraint or structuring should be applied (Dillenbourg, 2002), how support might be faded over time (Wecker & Fischer, 2007) or offered in a just-in-time fashion (Gweon, Rosé, Zaiss, & Carey, 2006), and what modality should be used for training prior to interaction, prompting during interaction, or eliciting reflection after interaction (Rummel, Spada, & Hauser, 2006). Use of process analysis for evaluation of supportive interventions is well established in the CSCL literature, see for example the extensive work by Weinberger and Fischer (Weinberger & Fischer, 2006). Process analysis approaches that are motivated by theories from Sociolinguistics can be argued to provide certain advantages in terms of providing insights that are cross-cutting with respect to theories of collaborative learning or learning more generally (Howley, Mayfield, & Rosé, 2013). In this chapter we employ a quantitative discourse analysis methodology informed by the Sociolinguistics literature called the Souflé framework (Howley & Rosé, 2011; Howley et al., 2013) paired with time series visualization using the Tatiana tool (Dyke et al., 2012) to pinpoint specific problems with a conversational agent based intervention at an early stage of development and inform its iterative development for effective support of collaborative interactions.

Conversational agents have a long history of successful support for individual learning with technology (Rosé, Jordan, Ringenberg, Siler, VanLehn, & Weinstein, 2001; Rosé & Van Lehn, 2005). A series of results offer hope that they can be used productively to offer support for collaborative learning, especially in chat environments (Kumar & Rosé, in press; Kumar & Rosé, 2011). The Cell Model study dataset we analyze in this chapter represents an early stage of an effort to develop a new style of Conversational agent-based support for collaborative learning, which can be viewed as structuring interaction both at the macro level (in all conditions), and at the micro level (in the Direct Agent and Indirect Agent support conditions).

In the remainder of the chapter, we first present an analysis of the main outcomes of the study in a style that is typical of quantitative analyses of experimental studies, which motivates a redesign, but does not offer insights into how the redesign should be done. This includes one cognitively focused dimension of the Souflé framework referred to as Transactivity, which we use as our main success metric at the level of collaborative process analysis. Next we present the two Sociolinguistic oriented dimensions of the Souflé analysis scheme, illustrated with examples from the data, in order to show what kinds of insights into social positioning within interaction can be revealed through this style of analysis. Finally, we present a Souflé style analysis of the full dataset as viewed through time series visualizations and draw hypothesized directions from it for redesign.

Main Analysis Motivating Redesign

Effect of Condition on Chat Discussion

In the Cell Model study, the goal of the experimental manipulation was to identify what type of support would most effectively elicit Academically Productive Talk (APT) from students, and subsequently, transactive contributions from their partner students, and finally, more learning. We begin by examining the conversation logs for evidence that we successfully manipulated prevalence of APT. Table 1 provides a summary.

The definition of APT moves is presented in the Cell Model study data chapter (Dyke, Howley, Adamson, Kumar, & Rosé, this volume), where we have listed five separate moves, namely Revoicing, Asking for a Rephrase, Asking for Agreement or Disagreement, Prompting for Elaboration, or Prompting for Justification. Transactivity is defined in two steps within the Souflé framework. First, a trasactive utterance must be an explicit display of reasoning, where a display of reasoning is one in which some evidence of a causal connection or compare/contrast relation is articulated. This display of reasoning is transactive if it evaluates or builds on a display of reasoning expressed earlier in the discussion. Here we adopt this one dimension of Souflé as a success metric for the manipulation.

Table 1 Average number of contributions (and percentage of contributions) of specific types across the three conditions of the study.

Condition	Academically Productive Talk Moves	Academically Productive Talk Moves from Any Source	Reasoning Moves	Transactive Moves
Unsupported	.56 (2.7%)	1.6 (1.8%)	1.6 (11%)	.55 (2.7%)
Indirect Agent	1.2 (4.9%)	3.8 (3.6%)	.53 (3.8%)	.13 (1.1%)
Direct Agent	.67 (6.4%)	4.25 (7%)	2 (17%)	.92 (5.1%)

The hypothesis that motivated the Cell Model study was that the agent based support would increase the prevalence of APT facilitation moves, which would then elicit more displays of reasoning and more transactivity. Thus, we begin our analysis by investigating the effect of condition on prevalence of APT, which may come either from the agent or from the students. There was no significant effect of condition on total number of APT moves contributed by students. However, there was a significantly higher total number of APT moves in both the supported conditions than the Unsupported condition when we count tutor contributions $F(2,42) = 5.5$, $p < .01$. And when we consider percentage of total contributions that are APT moves, we find a significantly higher percentage in the Direct agent condition than either of the other two conditions $F(2,42) = 13.9$, $p < .0001$. Thus, the first part of the hypothesis was partly supported.

The biggest difference between conditions shows up in terms of explicit displays of reasoning. Here there is a marginal main effect of condition on total number of reasoning moves per class period $F(2,42) = 2.46$, $p < .1$, whereby a student-t posthoc analysis demonstrates that students in the Direct condition produce a significantly greater number of reasoning moves than students in the Indirect condition, with the Unsupported condition not being significantly different from either. There was also a significant effect on percentage of reasoning moves $F(2,42) = 4.47$, $p < .05$, again where students in the Direct condition produce a significantly greater number of reasoning moves than students in the Indirect condition, with the Unsupported condition not being significantly different from either. Articulation of reasoning is a precondition for transactivity, so the second part of the hypothesis is also partly supported.

Nevertheless, while the effect on displays of reasoning is promising, the hypothesis begins to break down at the second step, in that we did not find evidence of any statistical relationship between the number or percentage of APT moves from the tutor and either student reasoning displays or transactive moves. We did, however, see a significant but weak correlation between total percentage of APT moves in a chat transcript from any source and the percentage of student contributions that were explicit displays of reasoning $R^2 = .11$, $p < .05$. This analysis shows that students in the Direct Agent condition were exposed to a greater percentage of contributions that were APT moves and that this higher percentage was associated with a greater percentage of explicit displays of reasoning during the chat. From this we have some weak indication that within the Direct Agent condition specifically, the manipulation was partly working as expected. The story was less positive in the Indirect condition. While the trend was for students in the Indirect Agent condition to employ a greater number of APT moves, the difference was not significant, and when we considered the trend in terms of percentage of contributions, it was not even the highest average. Furthermore, there was no significant main effect of condition on transactivity either within the Direct Agent condition or the Indirect Agent condition.

Overall, we find evidence suggesting some positive effect of condition on prevalence of APT, which is associated with increases in displays of reasoning. However, the goal of increasing transactivity was not reached. The big question here is why there are significantly fewer explicit displays of reasoning in the Indirect Agent condition when across conditions there is a significant positive correlation between APT moves and reasoning moves, and the trend is for there to be more APT moves in the Indirect Agent condition and the other conditions. This pattern suggests that something problematic was occurring, and thus this question will drive much of our process analysis in the subsequent sections of the chapter. In finding an explanation for this pattern, we arrive at insights that ultimately informed our redesign and lead to much more successful implementations of APT agents in subsequent studies (Dyke, Adamson, Howley, & Rosé, under review; Adamson, Jang, Dyke, & Rosé, under review; Adamson, Ashe, Jang, Yaron, & Rosé 2013).

Effect of Condition on the Whole Class Discussion

In addition to having an effect on behavior within the chat discussions, we also hoped that the collaborative exercise would serve to prepare students for the whole class discussion that took place on the day after the collaborative exercise and before the post-test. A summary of student contribution to the whole class discussion that followed in the next class period after the online activity is displayed in Table 2.

Table 2 Average number of contributions to the whole class discussions for students within class periods from each condition.

Period	**No Support**	**Indirect Agent**	**Direct Agent**
Period 1	4.2 (3.7)	8.0 (5.9)	3.7 (2.1)
Period 3	N/A	19 (8.5)	60 (49.5)
Period 6	1 (0)	3.2 (2.1)	5.8 (5.3)
Period 9	1 (0)	20 (0)	7 (0)

Analysis of two years of recorded classroom discussions from the whole study demonstrates a pattern whereby teacher adoption of APT practices showed a gentle increase over time that was punctuated by local increases in the discussions that immediately followed online CSCL activities (Clarke, Chen, Stainton, Katz, Greeno, Resnick, Howley, Adamson, & Rosé, 2013). This provides more substantial evidence of a connection between the online activities and the whole class discussions that followed them than what we could hope to see from simply examining the one discussion that was part of the Cell Model study. Thus, here we limit ourselves to an informal analysis.

For this chapter, we simply did a cursory analysis of differences in student behavior between conditions in the context of the whole group class discussion that was part of this Cell Model study only. The goal was to investigate the extent to which students in the supported conditions behaved differently than students in the unsupported condition. In this cursory analysis, we simply measured how active the students were in the discussion in terms of number of contributions to the discussion. Because the data were far from normally distributed, we first did a log transformation on the counts of contributions. We then performed an ANOVA analysis to determine whether there was a significant effect of condition. Since there was also a big difference in participation across class periods, we retained class period as an additional factor in the ANOVA analysis. Both class period ($F(3,21) = 7.0$, $p < .005$) and condition ($F(2,26) = 4.2$, $p < .05$) were statistically significant. A student-t posthoc analysis demonstrated that students in both the Direct Agent and Indirect Agent condition contributed to the whole group discussion significantly more frequently than students who had been in the Unsupported condition. In both cases the effect size was about .75 standard deviations. Thus, we have some evidence that the supported conditions better prepared students for active participation in the group discussion.

In addition to the effect on student activity during the whole class discussion, there was also a significant positive correlation between percentage of APT moves from the chat and the log of number of contributions to the whole class discussion. However, as a caveat we must mention that there were many student entries in the whole class discussion where we were not able to uniquely identify the student from the recordings, transcriptions, and notes. We do not have a reason to suspect that our ability to identify which students were contributing to the whole group discussion was biased by condition. However, we cannot completely eliminate the possibility that our analysis doesn't accurately reflect the effect of condition on participation. Overall, we consider the evidence of a connection between the experimental manipulation in the CSCL activity and behavior in the whole group discussion in the Cell Model study specifically to be merely suggestive. Despite this small glimmer of hope, based on this analysis, the main conclusion is that manipulation was not successful overall in eliciting the kind of behaviors we wanted from students. This was even clearer when we considered the pattern of results on learning gains.

Effect of Condition on Learning

Perhaps the most disappointing summative result from the Cell Model study was the lack of effect on learning overall. As mentioned in the data chapter (Dyke et al., this volume), domain knowledge was measured at three time points using a paper based test that included both multiple choice questions and open ended, explanation type questions. In order to capture how well the students grasped the material based on their test answers, we differentiated ability to remember what was observed (weight changes, indicator color changes), to understand what the observation meant (weight change means water movement, indicator color change means glucose and starch presence), and to explain and generalize (glucose always moving down the concentration gradient towards equilibrium, starch unable to move through the semi-permeable membrane). As the questions asked did not specifically request students to remember or understand, but rather to explain, the students with the better understanding tended to obtain low scores on the first two measurements and better on the final one. As a holistic measure of performance, we also ranked the explanation answers from "best" to "worst", using pairwise comparison, rather than a scoring rubric. Table 3 contains a summary of post-discussion test scores per assessment category across the Supported and Unsupported conditions in Period 1 only, which is the only class period where we observed significant pre to posttest learning gains.

To investigate the effect of condition on learning between the pretest and the postdiscussion test, we first verified that the random assignment of students to conditions was successful. This was confirmed in that there was no significant difference between conditions on the pretest $F(2,46) = .39$, $p = .69$. Next we checked whether there was evidence of learning. Overall, there was no general effect of learning across conditions. There was a significant effect of Class Period on all post-test measures, which showed that only students in the first class period showed significant pre to post test learning gains. Thus, we include an indicator of Class period as a fixed effect in our learning analysis. Due to low statistical power, we conduct a simplified analysis on learning measures, dividing students into Supported (which includes both the Direct and Indirect conditions) versus Unsupported, as well as Class Period1 versus Other Class Periods.

Table 3 Average post test score per assessment category in the Supported versus Unsupported conditions in Period 1 only, which was the only period that showed any significant pre to post test learning gains.

	Supported (Period 1)	**Unsupported (Period 1)**
Remember (High is better)	.6 (.54)	.66 (1)
Understand (High is better)	.4 (.54)	.89 (1.1)
Explain (High is better)	2 (.7)	1.1 (.9)
Experiment (High is better)	2.2 (1.3)	1.6 (1.4)

Ranking (Low is better)	2.4 (1.1)	4.8 (2.6)

The only significant effect on learning was on the Explain portion of the test, where there is a significant interaction between the two factors: $F(1,46) = 4.3$, $p < .05$, whereby within the first class period, students scored significantly higher in the Supported condition than the Unsupported condition with an effect size of 1.1 standard deviations. For the other assessments, the trend is in favor of unsupported conditions for recalling details of the experiment and demonstrating understanding of how indicators work in the experiments. However, since students tended to either focus on explanations or these more shallow measures, we consider that lower scores here might actually be preferred. Furthermore, for all other assessments, particularly those that demonstrate understanding of the domain concepts relating to diffusion and explanation for the results observed, the trend is in favor of the Supported conditions. Overall, the trend is weakly in favor of the Supported conditions over the Unsupported condition, although the effect is not strong nor clear.

Considering the whole set of results, the Direct agent condition shows the most promise, whereas the Indirect condition shows the least promise. Overall we must conclude that a redesign is in order. Consistent with the result on the online chat, the Indirect Agent condition shows up as the most problematic. What this analysis does not tell us is what change in design of agent behavior is likely to be more successful at achieving our goals. For this, we turn to the Souflé analysis scheme.

Souflé Analysis Scheme

In order to gain more insights into the inner workings of the chat that occurred during the online activity, we will use the Souflé analysis scheme (Howley & Rosé, 2011; Howley, Mayfield, & Rosé, 2013) to obtain a linguistic analysis of the discussion data from the Cell Model study. Souflé has three primary dimensions, one of which is cognitively focused, which we refer to as Transactivity, and the other two of which are more socially orientated, which we refer to as Authoritativeness (Martin & Rose, 2007, chapter 7) and Heteroglossia (Martin & White, 2005, chapter 3). Together these dimensions can be considered a multi-dimensional way of viewing leadership taking in conversation. While the cognitive dimension of that analysis scheme, which codes contributions according to whether they explicitly display reasoning, and whether that reasoning represents a new direction in reasoning within the conversation or builds on previously displayed reasoning (Sionti et al., 2011; Gweon et al., in press), other dimensions of that scheme are drawn from the broad subfield of linguistics referred to as Systemic Functional Linguistics (SFL). We discussed the transactivity dimension earlier in the chapter. Thus, we focus here on the two more socially oriented dimensions. An alternative presentation of the Souflé framework is presented in the PLTL Chemistry section of this volume (Howley et al., this volume).

The Authoritativeness and Heteroglossia dimensions of Souflé were inspired by work in the field of Systemic Functional Linguistics, which is a largely descriptive linguistic tradition that provides a firm foundation in analyses of genres of writing or text-based interaction (Martin & Rose, 2007; Martin & White, 2005), as well as face-to-face interaction (Veel, 1999).

Let us first consider the issue of authoritativeness in presentation of self. This analysis is derived from an operationalization of Martin and Rose's negotiation framework (Martin & Rose, 2003). In the negotiation framework, authoritativeness is demonstrated by making a contribution to a discourse that is not offered as an invitation for validation from another group member. For example, an assertion that is made in response to a question that is framed as a hint rather than a serious question, and then followed by an evaluation, is not coded as an authoritative assertion. The Negotiation Framework is a measure of authoritativeness where authority is demonstrated by making a contribution to a conversation that is not meant to be validated by another group member. In our operationalization of the Negotiation framework, there are four core moves, and two secondary moves:

- K1 (Primary Knower), in which the speaker considers herself to be the primary authority on a given (expressed) piece of knowledge
- K2 (Secondary Knower), when the contributor asks for knowledge from someone of higher authority
- A1 (Primary Actor), for contributions that display that one can perform a particular action.
- A2 (Secondary Actor), when instructing someone else to do an action, allowing the other person to either perform the action or reject the request.
- ch (Challenge), in which a speaker rejects the authority of the previous speaker to make the previous move
- o (Other), which encapsulates all other moves that do not fit in the five described above

For our purposes, "Primary Knower" and "Secondary Actor" moves are considered more authoritative (with respect to social relationships), while "Secondary Knower" and "Primary Actor" moves display less authoritativeness. In this chapter we are primarily concerned with authoritativeness over knowledge. As such, to compute a meaningful ratio for the authoritative moves, the formula would be: K1 / (K1 + K2).

What we can see here is that we can use the Authoritativeness ratio to pick out students that when we look more closely at the content of their interactions with others either appear to be having trouble (in the case of low authoritativeness) or look like leaders (in the case of high authoritativeness). This affirms the face validity of this coding. In other work, we have seen positive correlations between Authoritativeness and self-efficacy and between Authoritativeness and learning (Howley et al., 2011) or task engagement (Howley et al., 2013). We also see that we can discern something about the dynamic within the group when we look at a student's Authoritativeness ratio in relation to that of the other group members.

In the first interaction displayed in Figure 1 below, s008 has low authoritativeness in comparison to the partner student that we see. In the second interaction displayed in Figure 2 below we see that s002 has relatively high authoritativeness in relationship to the two partner students. In this Cell Model study dataset, we find that both Authoritativeness and difference between an individual's Authoritativeness and the average Authoritativeness of the other two students in the group both individually significantly correlate with the amount of reasoning a student contributes. This suggests that students with high Authoritativeness are students who treat themselves as sources of reasoning in the interaction. In order to facilitate transactive knowledge integration, we want all students within a group to treat themselves this way. Thus, we have reason to hope for groups where there is relative equality of Authoritativeness between students. However, as a caveat we must note that whereas an individual's Authoritativeness explains 11% of the variance, the difference in Authoritativeness between the individual and the group explains 15%. This suggests that there might be something more complex going on in the interplay between the students in terms of their Authoritativeness, and in understanding the impact of an intervention on an individual's behavior, we must first factor out the effect of the group dynamic, which also has a significant effect.

Speaker	Contribution	Negotiation
Alex(Tutor)	Now discuss what you observed in the video about conditions A and B. Compare what you observed to your predictions.	a2
s008	does ne one know wat they r doing	k2
	(1 minute pause)	
Alex(Tutor)	You should now move on to discussing what you observed in the video about condition C. Compare what you observed to your predictions. How is this different from what happened in condition A?	a2
	(2 minute pause)	
Alex(Tutor)	Ok, I gotta go.	o
Alex(Tutor)	It was nice talking to you all. :-)	o
	(3 minute pause)	
s008	ok this video made no sense at all soes any one know what they are doing	o
s012	maybeee	o
	(1 minute pause)	
s008	well then tell me	k2
s008	well then tell me	o
s012	well A is retarded	k1
s012	it gained twice then lost weight	k1
s008	well i only got the 1 hour thing nd the 5 hour one i dint watch the rest of the video	o

Figure 1 S008 is a student with low authoritativeness

Speaker	Contribution	Negotiation
s006	is anyone else seriously confused	k2
s010	{s006}, did you get any other observation other than the weight	k2
s010	yes	k1
s006	no i didnt write anything down for the observations	k1
s010	nice	o
s010	....	o
Alex(Tutor)	Ok, I gotta go.	o
Alex(Tutor)	It was nice talking to you all. :-)	o
s010	um bye?	o
s002	for 1 hour on a it was .620 they were both clear tube and beaker and	k1
s002	same for the B but the weight was .540	k1

Figure 2 S002 has high authoritativeness

The distinction between the core moves and other moves within the Negotiation framework allows us to find the more meaty portions of the interactions. The codes also allow us to see more when we examine them within different spans of the interaction than the statistics computed over the entire interaction. If we compute the statistics within segments of the conversation, we can identify particular segments where a speaker is behaving in a way that is uncharacteristic, for example, a relatively non-authoritative student taking an authoritative stance on some issue. We might refer to these as pivotal moments in the collaboration. We can also investigate whether that uncharacteristic behavior is related to some agent behavior, which may have a different effect locally than we see overall.

The Heteroglossia framework offers a complementary picture to the one just explored in relation to Authoritativeness. Within a heteroglossia analysis, assertions framed in such a way as to acknowledge that others may or may not agree, are identified as heteroglossic. The Heteroglossia framework is operationalized from Martin and White's theory of engagement (Martin & White, 2005), and here we describe it as identifying word choice that allows or restricts other possibilities and opinions. In other words, it captures the extent to which the students show openness to one another's views. This creates a rather simple divide in possible coding terms for contributions:

- Heteroglossic-Expand (HE) phrases tend to make allowances for alternative views and opinions (such as "She *claimed* that glucose will move through the semi-permeable membrane.")
- Heteroglossic-Contract (HC) phrases attempt to thwart other positions (such as "The experiment *demonstrated* that glucose will move through the semi-permeable membrane.")
- Monoglossic (M) phrases make no mention of other views and viewpoints (such as "Glucose will move through the semi-permeable membrane.")
- No Assertion (NA) expressions are ones that do not make an assertion, and therefore cannot be either monoglossic or heteroglossic.

We also code commands expressed as suggestions as Heteroglossic Expand, whereas ones that are stated simply as commands are coded as Monoglossic. Note that it is not always the case that heteroglossic assertions framed as negative polarity statements perform the function of contracting the set of options under negotiation. For example, if it is a constraint that is eliminated, then more items are made negotiable since fewer constraints need to be satisfied.

In looking at this analysis of contributions in terms of Heteroglossia, we see how a different attitude is communicated by use of these different styles. See Figure 3 for an example of a group with a highly monoglossic style.

Speaker	Contribution	Heteroglossia
Alex(Tutor)	You are now going to watch a video showing the cell in Conditions A, B and C.	m
Alex(Tutor)	As you watch the video, write down your observations on your worksheet.	m
s059	umm a you can say that when you mix them together the colors will change and something may happen..okay	he
Alex(Tutor)	Go to the Videos folder on the Desktop, and watch the video which is there.	m
s056	maybe	na
s062	wacth the video	m
s059	okay so do you want to put that down	he
s056	who is this anyway	na
	(3 minutes pause)	
Alex(Tutor)	Is everyone back?	na
s059	{s059}...and we can say when you mix them together the color changes and they have different weight	he
Alex(Tutor)	Now discuss what you observed in the video about conditions A and B. Compare what you observed to your predictions.	m
s062	glucose will dissolve in the distilled water	m

Figure 3 This chat segment displays a group with a highly monoglossic style.

As we see in Figure 3, the monoglossic style does not come across as welcoming or open to discussion. When we consider that, it is not so surprising that when our research group examined giving dialog agents heteroglossic, monoglossic, and neutral language in an idea generation task, we found that dialog agents with heteroglossic language result in the greatest idea generation productivity in a group task (Kumar et al., 2011). Results from that controlled comparison raise questions about whether students may respond to differences in Heteroglossia in tutor agent prompts that arise inadvertently when an intentional effort to control that aspect of style has not been made. We found, for example, in a posthoc analysis of tutor agent prompts across conditions that the proportion of Heteroglossic Expand contributions in the Indirect agent condition was significantly higher than that of the other two conditions $F(2,42) = 4.03$, $p < .05$. This appears to have been caused simply because the majority of tutor prompts were Monoglossic in style, but the prompts that were triggered to support Academically Productive Talk in the Indirect Agent condition were in the style of Heteroglossic Expand.

The final dimension in the Souflé framework is transactivity. It can be viewed as a means for displaying receptivity to the idea leadership of other students in the group. We have already discussed Transactivity in a previous section where we offer the summative analysis of study outcomes.

Together, the multi-dimensional Souflé framework allow us to view how idea leadership plays out in group knowledge construction and how it is possible to present one's views as standing on their own without denying others the right to have their own voice. But beyond this it might allow us to capture something important in the interplay between this leadership taking and an experimental manipulation. If the manipulation affects how students position themselves socially, it may also affect how students engage in idea contribution and idea integration within the interaction. It is precisely this sort of interplay that we begin to see when we apply the Souflé analysis to the whole corpus and view how it plays out over time using time series visualizations.

Souflé Style Analysis

Table 4 Overview of distribution of Souflé moves as well as Cheating and Offtask across conditions.

Condition	Authoritativeness Ratio	Percent Student Heteroglossia	Percent Student Reasoning	Percent Cheating	Percent Offtask
Indirect Agent	.54 (.21)	.08 (.06)	.03 (.1)	.18 (.15)	.35 (.15)
Direct Agent	.6 (.21)	.24 (.14)	.17 (.14)	.05 (.09)	.27 (.16)
No Support	.62 (.27)	.2 (.16)	.11 (.1)	.12 (.2)	.13 (.17)

Table 4 displays a summary of the Souflé style analysis, which includes average Authoritativeness ratio per condition as well as the average percentage of Heteroglossic Expand moves and explicit display of reasoning moves. Additionally, we added two informal diagnostic indicators, namely percentage of contributions that were off-task, which included places where students were joking around or insulting one another, and percentage of cheating moves, where students were engaged in passing answers back and forth rather than discussing the material. We do not find any significant effect of condition on percentage of cheating contributions, although there is trend for there to be the least cheating within the Direct Agent condition and most within the Indirect Agent condition. Similarly, there is significantly more off-task behavior in the two supported conditions than in the No Support condition $F(2,41) = 7.7$, $p < .005$, although the trend is for there to be more in the Indirect Agent condition than the Direct Agent condition. This adds further support for the conclusion that the Indirect Agent condition was problematic. But we are still left wondering why and what to do about it.

We already mentioned that there was a significant effect of condition on Percent Student Reasoning. As mentioned in the previous section, the big question here is why there are significantly fewer explicit displays of reasoning in the Indirect Agent condition when across conditions there is a significant positive correlation between Academically Productive Talk moves and Reasoning Moves, and the trend is for there to be more Academically Productive Talk moves in the Indirect Agent condition than in the other conditions. If we do a multiple regression with Total percentage of Academically Productive Talk moves, student

Authoritativeness, and the average Percent Heteroglossia of the other two students in a group, all of these factors have a positive significant correlation with Percent Student Reasoning, and together they explain 44% of the variance in Percent Student Reasoning. Without total percentage of Academically Productive Talk moves, we can only explain 38%. Thus, we have evidence that Academically Productive Talk moves positively contribute to the conversation. However, the condition that was meant to elicit Academically Productive Talk moves did not. We see from this that we have more evidence that we need to find a different way to elicit this behavior other than that the goal itself was not appropriate.

One clue is in connection with the Heteroglossia dimension. Overall, we find a positive and strong correlation between the average percentage of Heteroglossic Expand contributions in a discussion and the percentage of a student's contributions that are explicit reasoning displays, $R^2 = .5$, $p < .0001$. We see a corresponding negative correlation between percentage of Heteroglossic Expand contributions and percentage of Off-task contributions, $R^2 = 12$, $p < .05$ and a similar trend for less cheating when the percentage of Heteroglossic Expand utterances are high. What we notice when we consider all of the dimensions of Souflé is that in addition to a significant drop in percentage of contributions that contain explicit reasoning displays, we also see a significantly smaller percentage of student contributions that are Heteroglossic Expand $F(2,41) = 6.79$, $p < .005$. A student-t posthoc analysis confirms that this percentage is significantly smaller in the Indirect Agent condition than either of the other two conditions. This is a surprise considering that a greater proportion of tutor agent moves are Heteroglossic Expand in the Indirect Agent condition, as discussed in the previous section. We mentioned in the previous section that a higher proportion of Monoglossic contributions is associated with a somewhat negative attitude. All of this connection between the Heteroglossia dimension and other variables supports this interpretation. Clearly something was happening in that condition that was disenfranchising students, but it wasn't simply that they did not want to respond to the Indirect Agent prompts, because if that were true, we would expect to see significantly fewer APT moves from students rather than a trend for more APT moves.

To get a better idea of what was happening, we constructed a visualization using Tatiana (Dyke et al., 2012) where we see the average distribution of codes on the Heteroglossia dimension across time within each of the three conditions. Since some off-task talk nevertheless received codes on the Heteroglossia dimension, but did not contribute towards the substantive part of the conversation, we marked these as off-task rather than any of the four heteroglossia codes. Tutor prompts are indicated with vertical bars. The black bars are the task-related microscripting prompts that were the same in all conditions. The first black bar is where the tutor asks the group to start making predictions. The fourth black bar is where the students are instructed to watch the video. The next black bar after that is when the tutor asks if the students returned. The last black bar is where the tutor says goodbye and leaves the group to interact on their own. The green bars represent places where the tutor intervened with a condition specific prompt in at least one of the conversations.

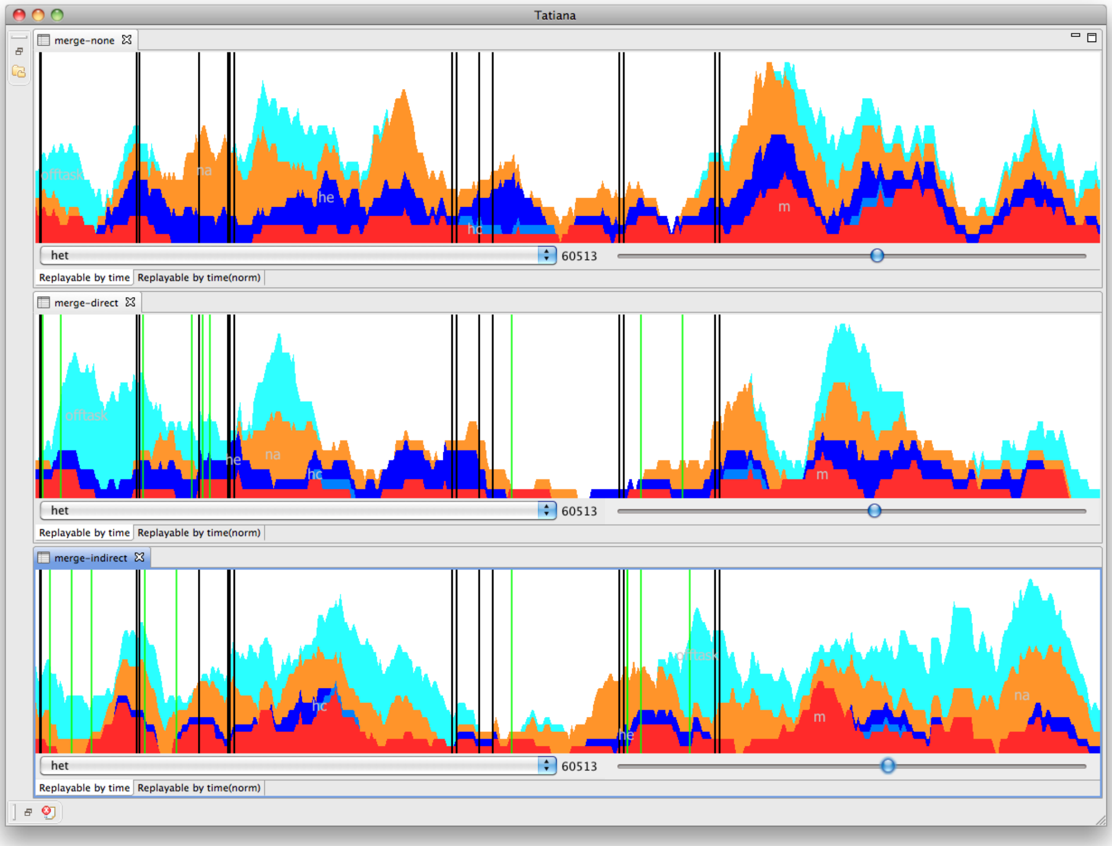

Figure 4 Tatiana visualization of Heteroglossia and Offtask codes over time.

The visualization tells an interesting story that can't be seen using average counts across whole discussions. First we notice that in all conditions, there is some off-task talk early and late in the period. However, in the Indirect condition, the initial segment of the conversation where there is intermittent off-task talk extends throughout the time when students were supposed to be watching the video, and starts up again before the tutor has left the conversation. In the No support condition, the initial off-task talk only extends halfway through the video, and doesn't start up again until after the tutor has left. In the Direct agent condition, students stop their off-task talk early in the period of time when they are making predictions and don't resume until after the tutor has left.

A similar interesting picture emerges when we examine patterns of monoglossic contributions over time. Despite the fact that the Indirect Agent condition was the highest on this, it is clear from the visualization over time that it did not start out that way. Monoglossia increases as more and more condition-specific prompts enter the conversation. The opposite is the case in the Direct Agent condition. Monoglossia starts out higher but is reduced as more of the condition specific prompts come in. In the unsupported condition, Monoglossia starts out at about the same level as in the Direct agent condition, but it does not reduce during the same period of time. The occurrence of Monoglossic contributions is relatively consistent within conditions, although distinct across conditions, between the time when the video is introduced and when the students were requested to return from the video. A substantial amount of heteroglossic discussion takes place in all but the Indirect Agent condition throughout that segment. After the students were instructed to return from the video, there is a period of intensive heteroglossic discussion both in the Unsupported condition and the

Direct Agent condition, but not in the Indirect agent condition, and it occurred earlier in the Direct Agent condition.

What we see is that in the Direct Agent condition, students respond to the agent by getting on-task and contributing in a heteroglossic style. Both of these show engagement and positive attitude. The opposite is the case in the Indirect Agent condition. The visualizations tell us that the answer should be found relatively early in the discussions after the agent has asked the students to start making predictions. Since we know that the students in the Indirect Agent condition did show some indication of attempting to follow the agent's request for them to engage in Academically Productive talk, the question is what happened when they did.

Alex(Tutor)	In condition C, we made a modification to condition A. We replace the glucose solution with a starch suspension. In order to detect the presence of starch, we replace the distilled water with Lugol's Iodine solution.	m
Alex(Tutor)	You should now move on to discussing what will happen in Condition C and your explanation for this change.	he
s041	the world is going to end in 2012	m
Alex(Tutor)	S027, now would be a good time to ask S034 to build on what S041 is saying.	he
s034	im so confused!	m
s027	034, would you like to build on to what 041 is saying? and me too!	he
Alex(Tutor)	When you are in agreement, write down your predictions and explanations for Conditions A, B and C on your worksheet.	m
s027	who is 34?	na
s034	{s034}	m

Figure 5 A frustrating interaction from the student perspective in the Indirect Agent condition

Figure 5 illustrates two things that likely contributed to the effect of the Indirect Agent prompts. Note first that responding to an Indirect Agent prompt requires more from students than responding to a Direct Agent prompt. For the latter, a student only has to make one comment, possibly building on or expanding on what another student has said, and then it's over. But in the Indirect Agent condition, first one student needs to respond with an Academically Productive talk move, and then another student has to respond to that. Thus, it takes both more time and more coordination. In both conditions, the tutor prompts were sometimes generated inappropriately as we see in Figure 5, however, in the Direct Agent condition, it was easier to disregard. Furthermore, we see an additional problem. Notice that s027 attempts to engage in Academically Productive Talk in response to the agent's inappropriate request. But then, to make matters worse, before the other student has time to respond, the tutor agent comes in with a task-related prompt, one that was scripted for a specific time in the interaction. Since these were given without regard for whether students had time to respond to the condition-specific prompts, the timed prompts were more interruptive in the Indirect Agent condition. Notice that after the task prompt, the students switched from their focus on the task, regardless of being confused, to off-task related talk.

The peaks in monoglossia in the Indirect condition were not always easy to explain, particularly for group C04 which did not experience any condition-related tutor prompts. For the other groups, these peaks were caused by four types of contributions: uncertainty about identities (e.g. "who is s011?" "I'm [s011], who are you"), expressions of confusion (e.g. "I don't understand this"), instructions (e.g. "write your predictions"), and giving answers ("condition [a] gained [weight]"). All four of these contribution types appear to have at their root the initial confusion about the task, exacerbated by the obliqueness of the indirect prompts which compound their lack of situatedness (asking students to perform Academically Productive Talk moves in inappropriate situations), the unfamiliarity of the situation (biology task, complex roles) and the problems over student identification. Distributing instructions and giving answers occur mainly when one of the students observes that his peers are struggling or confused and takes charge in resolving the situation. Lastly, the small heteroglossia peaks in the Indirect condition come from students following the tutor prompts and performing the requested Academically Productive Talk move, whether appropriate or not, increasing yet again student dissatisfaction and confusion.

Conclusions and Ideas for Redesign

One obvious problem with the condition-specific prompts was that they were sometimes triggered inappropriately, for example when students had uttered something in the form of a prediction that was actually off-task talk. However, this was the case both in the Direct Agent condition and the Indirect Agent condition. We do not see evidence that students responded the same way on average to the two kinds of prompts in the patterns of heteroglossia codes. Thus, it appears that a more serious issue was the lack of coordination between the condition-specific prompts and the timed task prompts, especially in the Indirect Agent condition where students were sometimes taxed by being interrupted while trying to follow the tutor's instructions. The realization of tutor interruptions conflicting with student actions is perhaps the most valuable insight that came out of the Souflé analysis, and provides the best direction for tutor design improvement.

From the summative statistical analysis, one might simply conclude that the Indirect Agent prompts were ineffective and annoying to students. However, a closer look with the Souflé analysis and time-based visualizations indicates that the real issue was a lack of coordination between types of prompts and a lack of consideration for how long it would take students to appropriately respond to the condition-specific prompts in the Indirect Agent condition. Considering the negative effect of the lack of coordination, it is no longer surprising that students showed a trend towards more APT in the Indirect Agent condition, but not a significant increase, and not the related increase we would have expected to see in explicit displays of reasoning that those moves are designed to elicit. Rather than drop the idea of the Indirect Agent condition, this analysis suggests that the timing and coordination problem are more important to address. This finding is consistent with recent work on triggering social prompts in agent supported CSCL environments where we have seen that trigger models that are trained using machine learning techniques on annotated data lead to significantly higher learning gains than other simpler trigger models (Kumar & Rosé, in press; Kumar et al., 2011).

The coordination issues revealed in this Souflé analysis were addressed in subsequent work on the Bazaar architecture for coordinating dynamic support for collaborative learning (Adamson & Rosé., 2012) and enabled subsequent positive results with APT agents (Dyke et al., under review; Adamson et al., under review; Adamson et al., 2013).

References

Adamson, D., Dyke, G., Jang, H., & Rosé, C. P. (Under Review). Towards Adapting Dynamic Collaboration Support to Support to Student Ability Level, submitted to the *International Journal of AI in Education*

Adamson, D., Ashe, C., Jang, H., Yaron, D., & Rosé, C. (2013). Intensification of Group Knowledge Exchange with Academically Productive Talk Agents. *Proceedings of the 10th International Conference on Computer Supported Collaborative Learning,* Madison Wisconsin, July 2013.

Adamson, D. & Rosé, C. P. (2012). Coordinating Multi-Dimensional Support in Conversational Agents, *ITS 2012 Proceedings of the 11th International conference on*

Intelligent Tutoring Systems, Lecture Notes in Computer Science Volume 7315, Springer-Verlag, pp 346-351.

Berkowitz, M., & Gibbs, J. (1983). Measuring the developmental features of moral discussion. *Merrill-Palmer Quarterly*, 29, pp 399-410.

Clarke, S., Chen, G., Stainton, K., Katz, S., Greeno, J., Resnick, L., Howley, H., Adamson, D., Rosé, C. P. (2013). The Impact of CSCL Beyond the Online Environment, *Proceedings of Computer Supported Collaborative Learning*

Dillenbourg, P. (2002). Over-scripting CSCL : The risks of blending collaborative learning with instructional design . In *Three worlds of CSCL - Can we support CSCL?* pp. 61–91 .

Dyke, G., Howley, I., Adamson, D., Kumar, R., Rosé, C. P. (this volume). Towards Academically Productive Talk Supported by Conversational Agents, in in Suthers, D., Lund, K., Rosé, C. P., Teplovs, C., Law, N. (Eds.). *Productive Multivocality in the Analysis of Group Interactions*, edited volume, Springer.

Dyke, G., Adamson, D., Howley, I., Rosé, C.P. (Under Review). Enhancing Scientific Reasoning and Explanation Skills with Conversational Agents, submitted to *IEEE Transactions on Learning Technologies*

Dyke, G., Kumar, R., Ai, H., Rosé, C. P. (2012). Challenging Assumptions: using sliding window visualizations to reveal time-based irregularities in CSCL processes, *in Proceedings of the International Conference of the Learning Sciences*. Sydney, Australia.

Gweon, G., Rosé, C. P., Zaiss, Z., & Carey, R. (2006). Providing support for adaptive scripting in an on-line collaborative learning environment. *CHI 06: ACM conference on human factors in computer systems*.

Gweon, G., Jain, M., Mc Donough, J., Raj, B., Rosé, C. P. (in press). Measuring Prevalence of Other-Oriented Transactive Contributions Using an Automated Measure of Speech Style Accommodation, *International Journal of Computer Supported Collaborative Learning*

Howley, I., Mayfield, E., Rosé, C. P. (2011). Missing Something? Authority in Collaborative Learning, in *Proceedings of Computer Supported Collaborative Learning*

Howley, I., Mayfield, E. & Rosé, C. P. (2013). Linguistic Analysis Methods for Studying Small Groups, in Cindy Hmelo-Silver, Angela O'Donnell, Carol Chan, & Clark Chin (Eds.) *International Handbook of Collaborative Learning*, Taylor and Francis, Inc.

Howley, I. & Rosé, C. P. (2011). Modeling the Rhetoric of Human-Computer Interaction, in *HCII'11 Proceedings of the 14th international conference on Human-computer interaction: interaction techniques and environments*, pp 341-350, Springer-Verlag Berlin, Heidelberg

Howley, I., Mayfield, E., Rosé, C. P., Strijbos, J. W. (this volume). A Multivocal Process Analysis of Social Positioning in Study Groups, in Suthers, D., Lund, K., Rosé, C. P., Teplovs, C., Law, N. (Eds.). *Productive Multivocality in the Analysis of Group Interactions*, edited volume, Springer.

Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämäläinen, R., Häkkinen, P., & Fischer, F. (2007). Specifying computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, 2(2), 211-224.

Kollar, I., Fischer, F., Hesse, F.W. (2006). Collaborative scripts - a conceptual analysis. *Educational Psychology Review* 18(2), 159–185.

Kumar, R. & Rosé, C. P. (2011). Architecture for building Conversational Agents that support Collaborative Learning, *IEEE Transactions on Learning Technologies special issue on*

Intelligent and Innovative Support Systems for Computer Supported Collaborative Learning

Kumar, R., Beuth, J., Rosé, C. P. (2011). Conversational Strategies that Support Idea Generation Productivity in Groups, in *Proceedings of Computer Supported Collaborative Learning*

Kumar, R. & Rosé, C. P. (in press). Triggering Effective Social Support for Online Groups. *ACM Transactions on Interactive Intelligent Systems*.

Martin, J. R., & Rose, D. (2007). *Working with Discourse: Meaning Beyond the Clause*. Continuum.

Martin, J. R., & White, P. R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., Weinstein, A. (2001). Interactive Conceptual Tutoring in Atlas- Andes, *Proceedings of AI in Education*.

Rosé C. P., & VanLehn, K. (2005). An Evaluation of a Hybrid Language Understanding Approach for Robust Selection of Tutoring Goals, *International Journal of AI in Education* 15(4).

Rummel, N., Spada, H., & Hauser, S. (2006). Learning to collaborate in a computer-mediated setting: Observing a model beats learning from being scripted. In: S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *Proceedings of the International Conference of the Learning Sciences* 2006 (pp. 634–640). Mahwah, NJ: Lawrence Erlbaum Associates.

Sionti, M., Ai, H., Rosé, C. P., Resnick, L. (2011). A Framework for Analyzing Develpoment of Argumentation through Classroom Discussions, in Niels Pinkwart & Bruce McClaren (Eds.) *Educational Technologies for Teaching Argumentation Skills*, Bentham Science.

Suthers, D. (2006). Technology affordances for inter-subjective meaning making: A research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning*, 1, pp. 315-337.

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaborations? In L. B. Resnick, C. Pontecorvo, & R. Saljo (Eds.), *Discourse, tools, and reasoning: Situated cognition and technologically supported environments*. Heidelberg, Germany: Springer-Verlag.

Veel, R. (2000). Language, knowledge and authority in school mathematics, in Christie, F. (Ed.) *Pedagogy and the Shaping of Consciousness: Linguistic and Social Processes*. Continuum

Wecker, C., Fischer, F. (2007). Fading scripts in computer-supported collaborative learning: The role of distributed monitoring . *CSCL'07 Proceedings of the 8th international conference on Computer supported Collaborative Learning*, 764–772 .

Weinberger, A., Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education* 46(1), pp 71-95.

Acknowledgments

This work was supported by the Pittsburgh Science of Learning Center and Graduate Training Grant from the Department of Education (#R305B040063).