Analysis of Human-to-Human Tutorial Dialogues: Insights for Teaching Analytics

Irene-Angelica Chounta, Bruce M. McLaren
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh PA, 15213, USA
{ichounta,bmclaren}@cs.cmu.edu

Patricia Albacete, Pamela Jordan, Sandra Katz
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, 15260, USA
{palbacet, pjordan, katz}@pitt.edu

ABSTRACT

In this paper we present a preliminary analysis of human-to-human tutorial dialogues as a precursor to developing an adaptive tutorial dialogue system, guided by a student model. One of our main goals is to further understand what makes tutorial dialogue successful, in particular how tutorial dialogues adapt to different student characteristics and prior knowledge and how to provide feedback to students in order to further support their practice. In particular we aim to identify important factors that affect tutorial dialogues and to characterize the level of support provided to students with different levels of understanding. Our approach and findings could also inform teaching and teaching analytics.

CCS Concepts

• Applied computing → Education.

Keywords

Tutorial dialogues; level of support; student modeling; adaptive dialogue; teaching; learning

1. INTRODUCTION

The learning process and its outcome highly depend on the social interaction between teachers and students and, in particular, on the proficient, helpful and focused use of language that students are exposed to by their teachers through written texts or tutoring discussions [20]. The findings from this research gave rise to several methodologies that promote dialogue as a means of keeping students engaged and motivated (such as *instructional conversation* [23]). This shift to socially-oriented methods was also observed for technology-enhanced learning contexts and for intelligent tutoring systems [26].

Our research goal is to develop an adaptive tutorial dialogue system, guided by a student model. In this paper we present a preliminary analysis of human-to-human tutorial dialogues as a precursor to achieving our main objective. Our approach and findings could also be informative to teaching and amenable to teaching analytics.

A key focus of our project is to further understand what makes tutorial dialogue successful; in particular, how tutorial dialogues adapt to different student characteristics and prior knowledge and how to provide feedback to students in order to further support their practice. According to Vygotsky, tutors use their assessment of students' ability to adapt the level of discussion to the student's "zone of proximal development" (ZPD)—that is, a little bit beyond the student's current level of understanding about a concept, ability to perform a skill, etc. [25]. In particular, we are interested in the following questions:

- How can we define the "level of support" that a successful tutor gives during tutoring?
- What makes some help given by a tutor more generous or stingy, easy or challenging, straightforward or "cognitively complex" for students?

Several researchers have addressed these questions, in various domains, for various purposes—for example, to develop instructional materials for classroom and computer-based learning environments; to address questions about how scaffolding and its counterpart, "contingent tutoring", take place in naturalistic learning settings in order to guide teacher training; to measure the effectiveness of curricula that implement scaffolding as a key feature. It is not surprising that the diverse set of goals driving the quest for ways to operationalize "levels of support" would produce an equally diverse set of descriptive frameworks.

The aim of this paper is thus threefold:

1. To explore how teachers regulate discussions and adapt their levels of support during tutorial dialogues,
2. To identify the factors that define "level of support" (LOS) in human-to-human tutoring examples, and
3. To propose analytics and mechanisms to guide tutors in orchestrating effective and efficient interventions in adaptive tutorial dialogues.

In the following sections, we present the process of identifying important factors and constructing a coding scheme to characterize the level of support in tutorial dialogues. However, our aim is not to use this scheme to analyze tutorial dialogues, but to provide guidance to dialogue authors for tailoring the level of support to provide to students who exhibit different levels of understanding.

2. RELATED WORK

Several researchers who have examined the question, "Why is human tutoring so effective?", have proposed that this effect is due to the highly interactive nature of human tutoring—in other words, the degree to which the student and tutor react to and build upon each other's dialogue moves and perceived understanding. This has been called the Interaction Hypothesis (e.g., [4,11,22]). However, an important line of research carried out over the past decade to test this hypothesis has found that it is not how much interaction takes place during tutoring that's important, nor the granularity of the interaction—for example, whether the student and tutor discuss a step towards solving a problem or the sub-steps that lead to that step. Rather, what matters is how well the interaction is carried out—for example, what content is addressed and how it is addressed, in a particular dialogue context (e.g., [5,6]).

In order to study and analyze the dynamics of dialogue, either in the classroom or in one-on-one settings, researchers have attempted to identify distinctive features of instructional dialogue and to define schemes for characterizing "level of support". Although most of the existing coding schemes were developed for problem-solving or other task-oriented domains, they may also prove relevant for operationalizing "level of support" for conceptually-oriented, reflective dialogues. The diverse set of goals driving the quest for ways to operationalize "levels of support" has produced an equally diverse set of descriptive frameworks. However, these schemes can be grouped according to the underlying, typically tacit dimensions that their developers used to differentiate the "levels of support" included in their coding schemes. The most common dimensions are the degree of detail (or specificity) in the tutor's help and the level of "cognitive complexity" in the tutor's comment, question, or directive. For example, Van de Pol's approach to measuring scaffolding involves characterizing the teacher's "level of control" [19]. The main dimensions in this scheme are the degree of "openness" or detail in the requested response, the length of the requested response and the amount of new content that the teacher introduced during her turn. Van de Pol proposed the measurement of "degree of teacher control" (*TDc*) on a six-step scale, starting from No Control (*TDc0* – when the teacher was not with the students) to Highest degree of control (*TDc5* – when the teacher provided new content, elicited no response and was providing the students with an explanation of the answer to a particular question).

Other schemes have focused on the distribution of cognitive effort between the tutor (also teacher or parent) and the student—in particular, who is doing the "heavy lifting" at particular points during instruction. Pino-Pasternak et al. [18] were interested in determining if the level of parental mediation impacted students' self-regulated learning (SRL) —that is, students' ability to control and monitor their own learning processes. They found that contingent shifting between mediation levels supported children's SRL. This scheme introduces the dimensions of cognitive demand (i.e. the distribution of cognitive effort between the parent and child), the student's level of understanding and the operationalization of "contingent tutoring" in terms of a mediation level that shifts to meet students' level of understanding. A similar approach was proposed by Nathan and Kim [16], who studied the way teachers regulate elicitations with respect to a cognitive hierarchy and in accordance with the correctness of students' responses. Toward that end, they coded teachers elicitations using Mehan's coding scheme [15].

In addition to the aforementioned schemes, Nystrand et.al [17] developed an approach to analyzing classroom discourse, focused on identifying factors that increase (or suppress) students' question-asking and other types of interactions that make up rich discussions, or "dialogic spells." Their approach includes a taxonomy that can be used to describe the cognitive complexity (which they call "cognitive level") of teachers' questions based on the level of abstraction and the status of information that the question invokes (i.e. new vs. old information).

Graesser et al. [10] focused on classifying the questions asked in tutorial dialogues. They defined 18 question categories based on their content. Furthermore, 8 of the aforementioned categories were further clustered into two subgroups: the questions that required a short answer ("*Verification," "Disjunctive," "Concept completion," "Feature Specification," "Quantification*") and the question that required long, elaborated responses ("*Definition," "Example," "Comparison").* The remaining categories were: *"Interpretation," "Causal antecedent," "Goal orientation," "Instrumental/procedural," "Enablement," "Judgmental," "Assertion,"* "*Request/directive*". Reasoning and deeper understanding are usually exposed with questions that ask "how" or "why" and invite for long, well-elaborated answers [10].

Scaffolding is a dynamic process. The tutor might change levels of support from one turn to the next and in accordance with the student's response. The main factor used to characterize the student's response to support is correctness: was the student's response to the tutor's question/hint correct, partially correct, or incorrect? Other factors that might influence change of the level of support have also been suggested, such as the difficulty level of the subject matter, time available and teachers' global perception of the student's ability (e.g. [7])

Human tutors are obviously unable to carry out detailed and highly accurate diagnoses of student knowledge [21]; their assessments of students' knowledge deficits are often inaccurate [3]. However, they nonetheless construct and dynamically update a normative mental representation of students' grasp of the domain content under discussion, as reflected in tutors' adaptive responses to students' need for scaffolding or remediation ([13]). For example, if a student solves a problem quickly and accurately, the tutor will probably challenge the student with some questions that go beyond the current problem's level of difficulty. On the other hand, if a student is struggling, the tutor will go slowly, perhaps clarifying step by step the knowledge the student seems to be lacking. As a tutoring session progresses, the tutor will dynamically update his or her conception of what the student knows and does not know. This allows the tutor to select appropriate problems to solve (*macro*–adaptation)—perhaps simpler problems if the student has not done well or more challenging ones if the student is performing proficiently. The tutor's dialogue with the student also enables the tutor to focus on particular curriculum elements (facts, concepts, skills, etc.) to discuss during a given problem and to determine the appropriate level at which to discuss these elements (*micro*-adaption). However, dynamically adjusting the level of support according to students' changing understanding of domain concepts is not a trivial task – for humans or intelligent tutors.

3. METHODOLOGY

3.1 Core Rationale

To better understand the mechanics of tutorial discussions, we studied a corpus of human-to-human tutorial dialogues. In particular, we took two approaches to better understand how tutors vary the "level of support" that they provide to students. The first was a fairly extensive literature review of coding schemes for tutoring discussions [7,10,11,16–19]. The second was to test the dimensions of level of support that other researchers have identified, by coding a corpus of tutorial dialogues. This was an iterative process of reviewing, coding and evaluating the results that resulted in the creation of a coding scheme for the level of support in tutorial dialogues.

3.2 Coding scheme

In order to test the accuracy and coverage of the factors that define "level of support" that we identified through our literature review, as summarized in the previous section, we coded tutors' turns in human tutor-student dialogues with respect to these factors (dimensions). In particular, we defined four dimensions:

- *Level of Control*: we have adapted the dimension of "Degree of Teacher Control" as described in [19] and coded it using a three-step scale (Low, Medium, High) where: *Low* control signifies that the teacher provided no new content and/or asked an open-ended question, expecting a well-elaborated answer; *Medium* control signifies provision of new content that is not directly related to the question or seeks a short answer; *High* control signifies that the teacher provides new content or provides a hint or elicits no response and instead provides an explanation. Table 1 presents three examples where the teacher provided feedback at different levels of control based on her perception with respect to the student's level of understanding of Newton's Second Law.

- *Question Category*: we have used the question categories as described in the coding scheme of Graesser et. al. [10]. In particular, we used 18 categories to code the teacher interventions, for example: verification, disjunctive, concept completion, verification etc. (presented in Related Work). Some examples on how this dimension was applied to the corpus are presented later on (see Table 7).

Table 1. Three examples of Low, Medium and High tutorial feedback regarding the Level of Control

	Level of Control		
	Low	**Medium**	**High**
Teacher	So what is the net force on her?		
Student	39N		
Teacher	***direction?***	***Which direction? Up or Down?***	***Which way? We have 500 N from the rope pulling up and 539N from her weight (the gravitational force from the earth) pulling down. So what is the direction of the net force?***

- *Level of Specificity*: this dimension refers to whether the tutor provided specific and focused information to the student. It was coded using a 3-step scale as: *Low* specificity signifies that the tutor does not provide detail or specific information to the student; *Medium* specificity signifies that the tutor provides some specific information related to the student's input but not enough to lead her to the answer; *High* specificity signifies that the teacher provides detailed feedback to the student, directly related to the issue in question. Specificity is an important factor of instructional dialogue and is usually perceived as an attribute of the information content that the teacher provides to the student (for example, Van de Pol differentiates between broad, open questions and detailed ones [19]).

In our case, this dimension refers to the specificity of the teacher's feedback in relation to the student input that precedes the tutor's turn and in this sense, it is different than the Level of Control. This means that a tutor turn could be coded as medium or high for Level of control and low for Level of Specificity. For example, let us consider the following dialogue:

> **Teacher:** *What minimum acceleration must the climber have in order for the rope not to break while she is rappelling down the cliff?*
>
> **Student:** the acceleration equals the rope tension divided by the climber's weight
>
> **Teacher:** *The general rule for finding acceleration is F= a*m. and this is known as Newton's Second Law of motion. But here your answer is not correct. Keep in mind that Newton's second law of motion can be formally stated as: "The acceleration of an object as produced by a net force is directly proportional to the magnitude of the net force, in the same direction as the net force, and inversely proportional to the mass of the object." So…?*
>
> **Student:** ….

In this example, the student provides a wrong answer on how to compute acceleration and the teacher replies with feedback regarding the general context. However the teacher does not provide feedback on why the student's answer is wrong or what in particular should be corrected.

Table 2. Three examples of Low, Medium and High tutorial feedback regarding the Level of Specificity

	Level of Support		
	Low	**Medium**	**High**
Student	S:F = ma		
Teacher	what's f there?		
Student	mg		
Teacher	***just mg?***	***just mg ? how many forces act on the climber ?***	***No.. the F in F=ma is always the net force on the object (or group of objects). The vector sum of all the forces on the object. I prefer to say "Sum of F= ma" because it's easier to get it right.***

Examples of the three levels of specificity are shown in Table 2.

- *Contingency*: to code contingency in the tutor's turn, we adapted the coding scheme of Pino-Pasternak et. al. [18]. According to this scheme, contingent tutoring takes place when the tutor challenges the student with questions or comments that are at or above her potential and non-contingent tutoring happens when the tutor poses questions and tasks that are lower than the student's potential. This dimension was coded on a binary scale (i.e. contingent vs. non-contingent). Examples of contingent and non-contingent tutorial feedback are presented in Table 3.

Table 3. Examples of continent and non-contingent teacher's input for the same Student-Teacher dialogue

	Contingent	**Non-Contingent**
Teacher	what is the acceleration then?	
Student	Fnet=ma, so a = 39/55	
Teacher	***Can you provide the units?***	***Ok, a = 39 N / 55 kg***

3.3 Dataset

We applied our coding scheme to part of a dialogue corpus that stems from previous research to assess the effectiveness of human-guided reflective discussions about physics problems [12]. In particular, the study involved sstudents who were taking an introductory physics course at the University of Pittsburgh; these students were randomly assigned to three conditions: one in which students received reflection questions and interacted with a human tutor via a chat interface; a second reflection condition in which students were asked the same set of reflection questions but received a static text explanation as feedback after they responded to these questions; and a third, a control condition in which students were not asked reflection questions but solved more problems than students in the other two conditions, in order to control for time on task. From this corpus, we chose three human-to-human tutorial dialogues on Newton's Second Law (i.e. "*The acceleration of an object as produced by a net force is directly proportional to the magnitude of the net force, in the same direction as the net force, and inversely proportional to the mass of the object*") from the first condition (students engaging in a typed dialogue with his or her tutor via a simple chat interface, about each reflection question) for further analysis.

The three dialogues were chosen to represent students who displayed three levels of gain from pretest to posttest: one low, one medium and one high. The problem and the reflection question for all three dialogues were stated as: "**Problem:** *A rock climber of mass 55 kg slips while scaling a vertical face. Fortunately, her carabiner holds and she is left hanging at the bottom of her safety line. Suppose the maximum tension that the rope can support is 500 N.* **Reflection Question:** *What minimum acceleration must the climber have in order for the rope not to break while she is rappelling down the cliff? (You do not have to come up with a numerical answer. Just solve for "a" without any substitution of numbers.)"*. In effect, this question asked students to name the forces that act on the climber and to apply Newton's Second Law in order to compute the acceleration.

3.4 Applying the coding scheme to our dataset

Four researchers (i.e. the authors of this paper – from now on referred to as "experts" for simplicity) were given an introduction to the coding scheme and the dialogues. They were also provided with a coding template and the rules and directions for how to code the dialogues. In particular, they were asked to code each tutor's turn for all three dialogues. An excerpt of one of the dialogues, for a "high gain" student, is shown in Table 4. The tutor's turn (highlighted in grey) was coded by experts with respect to four dimensions that relate to the Level of Support. Overall, the experts coded 19 tutor turns. When they completed the coding process, they participated in a focus group where they discussed their coding, the process of applying the dimensions, and problems or challenges that they faced in doing so. The results of the coding process and the comments and concerns of the experts are presented in the next section.

Table 4. Excerpt from a tutorial dialogue between a student (high gainer) and a tutor.

Student: 500/55 kg=a m/s^2
Teacher: I don't agree - that's the acceleration that just the pull from the rope would produce (well once the units are straightened out it would be). Think a little more
Student: I'm stuck. I know you have to take into account her weight and an additional acceleration to account for the extra 39N, but I'm not really sure how they fit together.
Teacher: All right. What is the general rule for finding acceleration from forces?
Student: F/m=a
Teacher: and what is the F there?
Student: tension?
Teacher: No.. the F in F=ma is always the net force on the object (or group of objects). The vector sum of all the forces on the object. I prefer to say "Sum of F= ma" because it's easier to get it right. So.. if she is sliding down and the rope is just short of breaking, what is the *net* force on her?
Student: 0
Teacher: hmm hmm that was what it was in the problem above. Now we are in the case where the rope breaks at >500N. What's the tension in the rope just short of breaking?
Student: 500N
Teacher: Right. that's pulling her which way?

4. Results

In order to assess the reliability of agreement between the experts, we computed the Fleiss' Kappa for all four coding dimensions of the scheme. Of course, these are only preliminary results and therefore the Fleiss' Kappa should be considered no more than a glimpse at the effectiveness and accuracy of the coding scheme. The results are displayed in Table 5. The inter-rater agreement can be interpreted as fair for the dimensions of Level of control and Question category and poor for the Level of specificity [2,14]. For the dimension of Contingency, the result is not statistically significant (p-value > 0.05). However, the inter-rater agreement results and the discussion that followed showed that the suggested coding scheme did not adequately capture the nuances of how tutors dynamically adapt their responses to student input, during human-to-human tutorial dialogue.

Table 5. Results of Fleiss' Kappa for the reliability of inter-rater agreement and for the four dimensions of the coding scheme.

Dimension	Fleiss' Kappa	p-value
Level of control	**0.404**	4.13e-11
Question category	**0.395**	0
Level of specificity	**0.141**	0.0245
Contingency	0.0764	0.415

We also gave the possibility to the coders to provide their comments or the reasoning for their codings while coding the dialogues. We further analyzed their free-text comments about their coding and additional explanations/justifications that they expressed during a focus-group discussion. Analysis of the free-text answers revealed that experts usually had different opinions about what the goal of the intervention was. In some cases, they even stated that most likely the teacher didn't have a specific goal but was instead trying to assess the student's knowledge state. Frequently, the experts stated that a specific intervention served multiple goals that related to both backward and forward functions. As backward, we define the part of the tutor's response that relates to the student's prior input and as forward, we define the part of the tutor's response that aims to provide hints, support, guidance to the student towards the correct answer [9]. A dialogue excerpt, along with one tutoring expert's comments about the teacher's turn, is presented in Table 6.

Table 6. Example of the expert's comments during the coding process with respect to backward/forward functions codes

Dialogue	Expert's comments
Student: 500/55 kg=a m/s^2	
Teacher: I don't agree - that's the acceleration that just the pull from the rope would produce (well once the units are straightened out it would be). Think a little more	First (in response to student answer): Show student that the answer is incorrect by telling him in what situation it would be correct. Second (to move forward): Given what the tutor said in "first" get the student to attempt to solve the problem again.
Student: I'm stuck. I know you have to take into account her weight and an additional acceleration to account for the extra 39N, but I'm not really sure how they fit together.	
Teacher: All right. What is the general rule for finding acceleration from forces?	First (in response to student answer): reassures student that it is okay. Second (to move forward): Get student to think about the correct answer by going to first principles.

Teachers often provided feedback and guidance in one dialogue turn. This caused mismatches in the coding of the four dimensions. For example, the experts stated that it was difficult to assess the Level of control of the teacher's intervention because she uses explicit hints to help the student but, at the same time, she poses an open-ended question. There were similar issues when assessing the Level of specificity. In that case, experts commented that it was hard because the teachers tended to give elaborated feedback on students' past answers but not further details on future steps. Therefore, it was not easy to decide on the specificity of the teacher's intervention. Finally, all experts agreed on the importance of these two dimensions, i.e. Level of control and Level of specificity but they expressed a need for more precise instructions for distinguishing between these categories, and applying them.

For the dimension of Question Category, the experts stated that the categories were too numerous (18 categories) and that in some cases multiple categories would apply to one intervention or that none of the existing categories was appropriate for some tutor turns. We present a related comment from an expert in Table 7.

Some experts also expressed doubts about how to apply the Contingency dimension. One mentioned that she coded a tutor turn as 'contingent' "*if the tutor's response reflected the level of understanding of the student and it mostly did*".

From the discussion that followed the coding process, it was obvious that the experts were not satisfied with the application of the coding scheme and the results. They stated that it was still unclear to them how teachers were effectively regulating the level of support during tutorial dialogues and how we can define metrics to provide support to teachers during the orchestration of dialogues. One of the main issues appeared to be the complexity of the dialogues themselves, as well as the ambiguity of both students' and teachers' interventions.

Table 7. Example of the expert's comments during the coding process with respect to question categorization

Dialogue	Question Category	Expert's comments
Student: 500/55 kg=a m/s^2		
Teacher: I don't agree - that's the acceleration that just the pull from the rope would produce (well once the units are straightened out it would be). Think a little more	Request/directive	It is like a request to "try again". Admittedly what the student is trying again is his/her previous attempt at quantification. Note the original question is not quantification - it doesn't ask for the value of a variable but the student interprets it that way.
Student: I'm stuck. I know you have to take into account her weight and an additional acceleration to account for the extra 39N, but I'm not really sure how they fit together.		
Teacher: All right. What is the general rule for finding acceleration from forces?	Definition	doesn't ask what does x mean but does the inverse (inverse of definition). No other (question category) seems to fit.

5. Discussion

In this paper we presented a preliminary study on the analysis of human-to-human tutorial dialogues. We carried out this analysis as a precursor to developing student modeling for an adaptive tutorial dialogue system. However, we believe that our approach could be informative to teaching and teaching analytics, especially for socio-oriented, constructivist approaches where dialogues between teachers and students are considered essential for the learning process.

The goal of this analysis was to review existing frameworks for coding and analyzing tutorial dialogues and to define a scheme for characterizing the level of support during dialogues—that is, how does a tutor effectively regulate when, and how much, support to provide? Four domain experts coded tutors' interventions in human-to-human tutorial dialogues and the results of the coding process were presented and discussed in focus group meetings.

From the results, it appears that the crucial factors in defining the levels of support are the amount of new content or new information shared by the tutor and the degree of detail (or specificity) in the tutor's help. The experts stated that it is very important to differentiate between the information that is offered as feedback to previous questions, the information that relates to the background knowledge that students may have and the information that is offered as hints or in order to push the student forward.

Even though the tutoring experts agreed that contingency between the teacher's and the student's turns plays an important role, the experts found it difficult to code student-tutor exchanges in terms of contingency. This may be related to the level of discourse analysis, which in this case was very low (i.e. at the exchange level, vs. at the episode and dialogue level). It is possible that we need larger sequences of tutor-student turns in order to appropriately assess contingency.

5.1 Towards a coding approach for characterizing "level of support"

So far, we studied the design and application of a coding scheme to define and characterize the Level of Support in tutorial dialogues. The coding scheme was designed based on a thorough literature review of related research. The results of applying the scheme revealed some weaknesses of the coding approach and the need for more precision in defining and applying some of its dimensions, as presented in the Results section. In light of these findings, we have further revised the proposed coding scheme.

The new coding scheme has four dimensions:

- **D1. Information related to the student's answer**: The first dimension refers to the amount and level of specificity of the information provided to the student and that is related to the student's prior answer. It is coded on a four-step scale (None, low, medium and high).
- **D2. Hints provision**: The second dimension refers to the hints that are provided to the student, either directly or through questions. It is coded on a four step scale (None, low, medium and high)
- **D3. Feedback on correctness**: This dimension refers to the feedback the tutor provides to the student's previous reply with respect to correctness. Attempts to move forward and ambiguous statements (i.e. "but what about the net force?") are not considered feedback. In the case where the tutor's turn does not follow a student's answer, this dimension is coded as non-applicable. This dimension is coded on a four step scale (None, Implicit, Explicit and Non-Applicable).
- **D4. Information related to the "Feedback on correctness"**: This dimension refers to the explanation or information the tutor provides about her feedback on the correctness of the student's previous turn. It is coded on a three-step scale (Yes – when the tutor provides an explanation along with feedback; No – when the tutor provides no explanation along with her feedback; Non-applicable).

Relative to the original coding scheme, we split the "Level of control" dimension into two: the "Information related to the answer" dimension and the "Hint provision" dimension. This was done because providing a hint is considerably different than providing information (sometimes the teacher just provides general information to describe the context) and thus, these two factors could not be captured by the same dimension. The dimension "Question Category" was eliminated because the coding results did not reveal solid relations between different categories and the level of support. Moreover, the experts mentioned that it was extremely complicated and hard to code the tutor's turn based on the list of categories we provided them. The dimension of "Level of Specificity" was also split into two categories: "Feedback on Correctness" and "Information related to 'Feedback on correctness'". Additionally, we eliminated the dimension of "Contingency" because, at least in this dialogue corpus, non-contingent tutor turns were rare.

Currently, we have only carried out trial applications of the coding scheme in order to refine the dimensions and the coding levels. However, the results so far are very encouraging both in terms of inter-rater agreement (Cohen's kappa for the four dimensions ranged from 0.764 to 0.871) and the experts' comments. Nonetheless, we need to further validate the coding scheme, by applying it to more data.

5.2 Limitations of the study

This study was part of a broader project that aims to enhance an adaptive tutorial dialogue system using student modeling techniques. Our goal was to characterize the various levels of support that teachers provide to students during human-to-human tutorial dialogues and to identify the factors that affect the provision of support. Towards that end, we have focused on characteristics of the tutors' feedback, such as the amount and specificity of information, the provision of feedback, etc. However, there are other factors that were not taken into consideration in this study. One important factor is the student model that the teacher mentally, dynamically builds and maintains for each student. The teacher builds this mental model of the student based on the student's answers. Based on this model, the teacher regulates the dialogue and the level of support, as the teacher deems appropriate. The effect that this might have on the teacher's feedback is demonstrated in Table 8 where we present a case from our corpus. Based on informal comments about students' ability level that this tutor expressed to one of the authors, we are aware that this tutor perceived student A as an underachiever and student B as an overachiever.

Based on the students' responses, both student A and student B do not understand the meaning of the net force. However, it is evident that the teacher provides more information and support to student B than student A.

Taking into consideration the effect the teacher's perception about students' overall ability level might have on the level of support is an extremely complicated issue that we have not addressed in our coding scheme. However, we acknowledge this is an important factor that should not be overlooked in developing more adaptive tutorial dialogue systems.

Table 8. Two examples of tutorial dialogues that reflect the tutor's perception of each student's overall ability

Student A (underachiever) - Teacher Dialogue	Student B (overachiever)- Teacher Dialogue
Student A: a = f / m	Student B: 500/55 kg=a m/s^2
T: what's f ?	*T: I don't agree - that's the acceleration that just the pull from the rope would produce (well once the units are straightened out it would be). Think a little more*
Student A: f = mg	Student B: I'm stuck. I know you have to take into account her weight and an additional acceleration to account for the extra 39N, but I'm not really sure how they fit together.

T: just mg ? how many forces act ont he climber ?	T: All right. What is the general rule for finding acceleration from forces?
Student A: mg + T	Student B: F/m=a
T: is mg down or up?	T: and what is the F there?
Student A: down and T is up	Student B: tension?
T: ok so now solve for a again plugging in T and mg	T: No.. the F in F=ma is always the net force on the object (or group of objects). The vector sum of all the forces on the object. I prefer to say "Sum of F= ma" because it's easier to get it right. So.. if she is sliding down and the rope is just short of breaking, what is the *net* force on her?
Student A: a = (mg + T) / m	Student B: 0
T: which direction is mg in ?	T: hmm hmm that was what it was in the problem above. Now we are in the case where the rope breaks at >500N. What's the tension in the rope just short of breaking?

Application of the coding scheme was carried out by the authors of this paper (“experts”). We plan to get input from domain experts (i.e. physics teachers) for our coding scheme and formally validate it further.

5.3 Dialogue-support mechanisms as teaching analytics

Our objective is to study the mechanisms driving human-to-human tutorial dialogue and use this information to create algorithms and principles to guide effective, automated tutorial dialogue use this information to create mechanisms and rules to support effective dialogue orchestration. In our case, we aim to enhance a dialogue-based intelligent tutor to support adaptive interactions. However, this line of research can be used to support teachers in other challenging settings, such as in large classrooms or in distance-learning scenarios, where the need for teaching analytics is prominent [8]. In particular, we envision the use of dialogue-related indicators to provide feedback to teachers and recommendations on how to appropriately support their students. This can be achieved by creating appropriate visualizations and data analytics based on dialogue-related indicators and integrating them into teacher dashboards. For example, we could provide visual indication of the amount of information a teacher provides to a student or a visualization of the content a teacher contributes to a topic in comparison to the content the student contributes to the same topic.

So far, teacher dashboards provide information about the tutor-student interactions that mostly has to do with the number of messages students exchange with the automated tutor, or the rate of exchange [24] (an exception to this is recent work by Aleven et al [1]) We can enhance this work by adding content-related or quality-related information, such as what concepts have been covered or how well students have elaborated on arguments. We could also recommend to teachers emphasizing certain aspects of the dialogue, such as, leaving time for student self-reflection or providing elaborated information instead of hints or feedback on correctness. This can be achieved by defining guidelines on feedback provision with respect to different student types and different levels of understanding. From our experience analyzing human-to-human tutorial dialogues, we came across several cases where the teacher would adapt the level of discussion based on her perception with respect to the student’s level of understanding, rather than the actual student’s response. It was evident that teachers provided more information and less hints to low achievers while they were reluctant to give away the answer or too much information to the high achievers.

For example, let us consider two students: Frank is a low performer who lacks basic knowledge in motion laws and who is not confident for his skills in physics. On the contrary, Nancy is a high performer with good background knowledge who enjoys studying physics. Their teacher has to provide appropriate feedback taking into account their prior knowledge and personal characteristics. Based on our observations of human-to-human tutorial dialogues, in the case of an incorrect student answer, the teacher might want to provide information and explanation to Frank, encouraging him to repeat basic concepts and definitions. For Nancy, the teacher would encourage her to try again and to check her line of reasoning for possible mistakes, without giving away the answer.

Defining explicit guidelines on what kind of feedback is appropriate for specific student types can assist the teacher in providing personalized student feedback. Furthermore, this set of guidelines can be helpful for students in teacher education and young professionals, who do not yet have the expertise to evaluate tutorial dialogues, especially in real time.

5.4 Future work

This paper presented a preliminary study of the work-in-progress on a project that aims to develop an adaptive dialogue tutoring system. Currently, we are working on the refinement of the coding scheme for the assessment of Level of Support for tutorial dialogues. So far, we have identified factors that affect the level of support in dialogues, focusing on computationally tractable dimensions; that is, dimensions that can be captured by automated or semi-automated measures and indicators, in order to develop an adaptive tutorial dialogue system.

Our primary focus in analyzing and coding “level of support” is to specify authoring principles for adaptive tutoring systems—that is, rules for how to tailor tutor responses for different levels of student understanding— with respect to a given domain, and with respect to specific domain knowledge components. Towards that end, we will also work with teachers. In future work, we will involve them in implementing a rule-based approach for structuring adaptive tutorial dialogues.

6. REFERENCES

1. Vincent Aleven, Franceska Xhakaj, Kenneth Holstein, and Bruce M. McLaren. 2016. Developing a Teacher Dashboard For Use with Intelligent Tutoring Systems (to appear).
2. Douglas G. Altman. 1990. *Practical statistics for medical research*. CRC press.
3. Michelene TH Chi, Stephanie A. Siler, and Heisawn Jeong. 2004. Can tutors monitor students’ understanding accurately? *Cognition and instruction* 22, 3: 363–387.
4. Michelene TH Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science* 25, 4: 471–533.
5. Min Chi, Kurt VanLehn, and Diane Litman. 2010. Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In *Intelligent Tutoring Systems*, 224–234.

6. Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. 2011. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education* 21, 1–2: 83–113.
7. Christine Chin. 2006. Classroom interaction in science: Teacher questioning and feedback to students' responses. *International journal of science education* 28, 11: 1315–1346.
8. Irene-Angelica Chounta and Nikolaos Avouris. 2014. Towards the real-time evaluation of collaborative activities: Integration of an automatic rater of collaboration quality in the classroom from the teacher's perspective. *Education and Information Technologies*: 1–21.
9. Mark G. Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, 28–35.
10. Arthur C. Graesser and Natalie K. Person. 1994. Question asking during tutoring. *American educational research journal* 31, 1: 104–137.
11. Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology* 9, 6: 495–522.
12. Sandra Katz and Patricia L. Albacete. 2013. A tutoring system that simulates the highly interactive nature of human tutoring. *Journal of Educational Psychology* 105, 4: 1126.
13. Sandra Katz, David Allbritton, and John Connelly. 2003. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education* 13, 1: 79–116.
14. J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*: 159–174.
15. Hugh Mehan. 1979. *Learning lessons*. Harvard University Press Cambridge, MA.
16. Mitchell J. Nathan and Suyeon Kim. 2009. Regulation of teacher elicitations in the mathematics classroom. *Cognition and Instruction* 27, 2: 91–120.
17. Martin Nystrand, Lawrence L. Wu, Adam Gamoran, Susie Zeiser, and Daniel A. Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes* 35, 2: 135–198.
18. Deborah Pino-Pasternak, David Whitebread, and Andrew Tolmie. 2010. A multidimensional analysis of parent–child interactions during academic tasks and their relationships with children's self-regulated learning. *Cognition and Instruction* 28, 3: 219–272.
19. Janneke Eva Pol and others. 2012. *Scaffolding in teacher-student interaction: exploring, measuring, promoting and evaluating scaffolding*.
20. Barbara Z. Presseisen and Alex Kozulin. 1992. Mediated Learning–The Contributions of Vygotsky and Feuerstein in Theory and Practice.
21. Ralph T. Putnam. 1987. Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American educational research journal* 24, 1: 13–48.
22. Carla van de Sande and James G. Greeno. 2010. A framing of instructional explanations: Let us explain with you. In *Instructional explanations in the disciplines*. Springer, 69–82.
23. Roland G. Tharp and Ronald Gallimore. 1991. The Instructional Conversation: Teaching and Learning in Social Activity. Research Report: 2.
24. Eleni Voyiatzaki and Nikolaos Avouris. 2014. Support for the teacher in technology-enhanced collaborative classroom. *Education and Information Technologies* 19, 1: 129–154.
25. Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children* 23, 3: 34–41.
26. Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.