Abductive Proofs as Models of Students' Reasoning about Qualitative Physics

Maxim Makatchev (maxim@pitt.edu), **Pamela W. Jordan (pjordan@pitt.edu)**,
Umarani Pappuswamy (umarani@pitt.edu) and **Kurt VanLehn (vanlehn@pitt.edu)**
Learning Research and Development Center
University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Abstract

In this paper we describe a part of the Why2-Atlas tutoring system that models students' reasoning in the domain of qualitative physics. The main goals of the model are (1) to evaluate correctness of the student's essay, and, in case the essay contains errors, (2) to direct remedial tutoring actions according to plausible errors in the student's reasoning. To meet these goals, a backchaining theorem prover generates a set of assumptions and a chain of reasoning (a proof) that plausibly led the student to write the observed essay. A proof can include correct as well as buggy reasoning steps and assumptions. After a proof is generated, it is analyzed for correctness and the analysis is used to generate appropriate feedback to the student. We describe the weighted abductive theorem proving framework, outline previous and upcoming evaluations and discuss possible future directions.

Introduction

The Why2-Atlas tutoring system is designed to encourage students to write their answers to qualitative physics problems as essays that include explanations of their arguments (VanLehn, Jordan, Rosé, Bhembe, Böttner, Gaydos, Makatchev, Pappuswamy, Ringenberg, Roque, Siler, & Srivastava, 2002). If the essay is incomplete or incorrect, the system generates elicitation or remediation feedback, respectively. For the purpose of evaluating completeness and correctness of the essay, a deep understanding of its contents is necessary. In the domain of qualitative physics, a deep understanding of the essay involves reasoning about logical relationships between the statements in the essay. The theorem proving approach that we use in Why2-Atlas provides the means to recover a structure of logical dependencies that connects the propositions representing (a) the essay text, (b) the problem statement, and (c) plausible student reasoning steps not explicitly stated in the essay.

Previously, formal methods for analyzing natural language text have encountered a number of challenges, such as the difficulty of obtaining propositional representations for sentences and the need for large amounts of commonsense knowledge in order to interpret the many concepts expressed. Consider, for example the qualitative physics problem presented in Figure 1 along with an actual student explanation. To address all the errors in the essay, the propositional representation of the essay must account for such commonsense knowledge as "An elevator has a ceiling" and "A human has a head."

Question: Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

Explanation: The keys are affected by gravity which pulls them to the elevator floor, because the keys then have a combined velocity of the freefall and the effect of gravity. If the elevator has enough speed the keys along with my head would be pressed against the ceiling of the elevator, because the acceleration of the elevator car along with me and the keys would overwhelm the gravitational pull.

Figure 1: The statement of the problem and an example explanation.

In addition, even if the system is able to represent the respective statements, it also needs to be able to reason about correctness of logical relations between these statements (a statement b can be false in the context of the problem, while the statement $a \rightarrow b$ can be true). Such reasoning would also be desirable for the purpose of providing a more substantive feedback to the student, e. g. excessive argumentation. An informal example of a possible chain of reasoning that student used to arrive at the statement "The keys would be pressed against the ceiling of the elevator" is shown in Figure 2.

An abductive theorem proving approach allows one to cope with propositions that cannot be proven due to lack of applicable rules by assuming such propositions are true without a proof (the operation is also referred to in the literature as *abducing*) when such assumptions allow a proof to be completed. The fewer the assumptions made, the better the proof. Weighted abduction takes this approach farther and assigns a cost to the set of assumptions depending on their individual weights and the chain of reasoning that led to the generation of these assumptions. A proof can include correct as well as buggy reasoning steps and assumptions.

In this paper we describe a part of the Why2-Atlas tutoring system that models students' reasoning of qualitative physics. The theorem prover, called Tacitus-lite+, is a derivative of SRI's Tacitus-lite (Hobbs, Stickel, Martin, & Edwards, 1988, p. 102) that, among other extensions, incorporates sorts. First we provide an overview of the knowledge representation. Next we describe the abduc-

Step #	Proposition	Justification
1	before the release, the keys have been in contact with the man, and the man has been in contact with the elevator	given
2	at the moment of release, velocity of the keys is equal to velocity of the elevator	bodies in contact over a time interval have same velocities
3	after the release, nothing is touching the keys	given
4	after the release, the keys are in freefall	if there is no any contact then the body is in freefall
5	after the release, the keys' acceleration is not equal to the elevator's acceleration	*elevator is not in freefall
6	after the release, the keys' velocity is not equal to the elevator's velocity	if initial velocity is the same and accelerations are different the final velocities are different
7	the keys touch the ceiling of the elevator	if the keys' velocity is smaller than the elevator's velocity, the keys touch the ceiling

Figure 2: An informal proof of the excerpt "The keys would be pressed against the ceiling of the elevator" (From the essay in Figure 1). The buggy assumption is preceded by an asterisk.

tive theorem proving framework and the heuristics we developed that aim at maximization of the plausibility of the proof as a model of the student's reasoning and the utility of the proof for the tutoring system. The measure of plausibility is evaluated with respect to (a) the misconceptions that were identified as present in the essay by the prover and by a human expert, and (b) the proof as a whole. The utility for the tutoring task can be interpreted in terms of relevance of the tutoring actions (triggered by the proof) to the student's essay, whether the proof was plausible or not. We also discuss the assumptions of cognitive economy and concept-level consistency that we make about the student in relation to the plausibility of the model. Next we summarize previous evaluations of the Why2-Atlas system (VanLehn et al., 2002) and of an early version of the abductive reasoning engine (Jordan, Makatchev, & VanLehn, 2003). Finally we conclude with a section on our future work.

Knowledge Representation for Students' Reasoning about Qualitative Physics

Envisionment and idealization

Generating an internal (mental) representation plays a key role for both novice and expert problem solving (Ploetzner, Fehse, Kneser, & Spada, 1999; Reimann & Chi, 1989). (Reimann & Chi, 1989) describes the internal representation in terms of "objects, operators, and constraints, as well as initial and final states." This notion of internal representation overlaps with *envisionment*, which is defined in qualitative physics problem solving (de Kleer, 1990) as a sequence of events described in the problem or implied by the description. A further step, translating the envisionment into the domain terminology (bodies, forces, motion properties) is referred to as *idealization* in (Makatchev, Jordan, & VanLehn, 2004a).

For the problem in Figure 1, for example, a possible envisionment is: (1) the man is holding the keys (elevator is falling); (2) the man releases the keys; (3) the keys move up with respect to the man and hit the ceiling of the elevator. The idealization would be:

Bodies: Keys, Man, Elevator, Earth.

Forces: Gravity, Man holding keys

Motion: Keys' downward velocity is smaller than the downward velocity of the man and the elevator.

Many misconceptions that students have are rooted in the envisionment and idealization (Ploetzner et al., 1999). To make the task of representing possible correct and erroneous envisionments feasible we restrict ourselves to problems with few plausible envisionments. The rules of mechanics, which rely mostly on the formal domain terminology, are augmented by rules for reasoning about most common envisionments, which use a looser language. Further we briefly describe these representations.

Qualitative mechanics ontology

The ontology for the subset of qualitative mechanics that the system addresses consists of bodies (e. g., keys, man), agents (air), phenomena (e. g., gravity, friction), and conventional physical quantities (e. g., force, velocity, position). To adequately represent justifications, we also have representations for physics laws (Newton's First Law) and basic algebraic expressions ($F = ma$). While internally the reasoning is done within a coordinate system that is fixed for each problem (for example, horizontal axis x directed to the right and vertical axis y directed up), a student's reasoning can be independent of coordinate system choice, operating instead in relative terms (up, down, in front of). The representation and corresponding translation rules are described in the following section.

Logical constants and variables, corresponding to bodies, agents, and quantities are associated with a sort symbol. Sorts are partially ordered by a natural subset order. Domains of the predicate symbols are restricted to certain sorts (so that each argument position has a corresponding sort symbol). These associations and constraints constitute an *order-sorted signature* (Walther, 1987).

Description	Sort
quantity	`Quantity1b`
identifier	`Id`
body (or two bodies in case of `force`)	`Body`
axial component or not	`Comp`
qualitative derivative of the magnitude	`D-mag`
quantitative derivative of the magnitude	`D-mag-num`
zero or non-zero magnitude	`Mag-zero`
quantitative magnitude	`Mag-num`
sign for axial component	`Dir`
quantitative direction	`Dir-num`
qualitative derivative of the direction	`D-dir`
beginning of time interval	`Time`
end of time interval	`Time`

Table 1: Slots of a vector quantity of sort `Quantity1b`.

Time is represented using time instants as basic primitives. Time intervals are denoted as a pair (t_i, t_j) of instants. This and the order relation `before` on time points enables us to reason about a reasonably rich subset of the mechanics domain.

Argument slots and an order-sorted signature for a predicate representing a vector quantity that involves a single body (for example `velocity`, `total-force`) are shown in Table 1.

A number of relation predicates are used to specify various algebraic and logical relations between physical quantities (see Table 2).

Two bodies can also be related via a state of `contact` with possible fillers `detached`, `attached`, and `moving-contact` (for the case of relative motion between bodies in contact).

Rules

The rules cover correct reasoning at the formal domain level of an idealized problem ("zero acceleration implies constant velocity"), buggy reasoning at this same level ("zero force implies decreasing velocity"), and some common relevant aspects of the idealization and envisionment stages ("if axis y is directed upward and velocity is vertical and positive then velocity is upward").

The rules are represented as *extended Horn clauses*, namely the head of the rule is an atom or a conjunction of multiple atoms. Further details of the knowledge representation are covered in (Makatchev et al., 2004a).

Weighted Abductive Theorem Proving

Order-sorted abductive logic programming framework

Similar to (Kakas, Kowalski, & Toni, 1998) we define the *abductive logic programming framework* as a triple $\langle T, A, I\rangle$, where T is the set of *givens* and *rules*, A is the set of abducible atoms (potential hypotheses) and I is a set of integrity constraints. Then an *abductive explanation* of a given set of sentences G (goals) is (a) a subset Δ of abducibles A such that $T \cup \Delta \vdash G$ and $T \cup \Delta$ satisfies I, and (b) the corresponding *proof* of G. The set Δ is *assumptions* that explain the goals G. Since an abductive explanation is generally not unique, various criteria can be considered for choosing the most suitable explanation (see Section "Proof search heuristics").

An *order-sorted abductive logic programming framework* $\langle T', A', I'\rangle$ is an abductive logic programming framework with all atoms augmented with the sorts of their argument terms (so that they are sorted atoms) (Makatchev et al., 2004a). Assume the following notation: a *sorted atom* is of the form $p(x_1, \ldots, x_n) : (\tau_1, \ldots, \tau_n)$, where the term x_i is of the sort τ_i. Then, in terms of unsorted predicate logic, formula $\exists x\ p(x) : (\tau)$ can be written as $\exists x\ p(x) \wedge \tau(x)$. For our domain we restrict the sort hierarchy to a tree structure that is naturally imposed by set semantics and that has the property $\exists x\ \tau_i(x) \wedge \tau_j(x) \rightarrow (\tau_i \preccurlyeq \tau_j) \vee (\tau_j \preccurlyeq \tau_i)$ where $\tau_i \preccurlyeq \tau_j$ is equivalent to $\forall x\ \tau_i(x) \rightarrow \tau_j(x)$.

Tacitus-lite+ uses backward chaining with the order-sorted version of modus ponens:

$$\frac{\begin{array}{l} q(x', z') : (\tau_5, \tau_6) \\ p(x, y) : (\tau_1, \tau_2) \leftarrow q(x, z) : (\tau_3, \tau_4) \\ \tau_5 \preccurlyeq \tau_3,\ \tau_6 \preccurlyeq \tau_4 \end{array}}{p(x', y') : (\min(\tau_5, \tau_1), \tau_2)}$$

Proof search heuristics

The aim of the proof search heuristics is to quickly find a proof that optimizes a measure of utility of the proof for tutoring applications and a measure of plausibility of the proof as a model of a student's reasoning. A highly plausible proof has a high value for its utility measure since it potentially allows the tutoring system to generate feedback that is more relevant to the student's actual mental state. However a less plausible proof would have the same utility measure if it results in the same tutoring action as a more plausible proof. In fact, we would prefer a less plausible proof over the more plausible proof, their utility measures being same, if the former takes less time to compute.

The plausibility measure is based on two cognitive assumptions. The first assumption, *cognitive economy*, can be interpreted in the context of the abductive proofs as a preference for a simpler proof structure (for example a smaller proof) and a smaller cost for the propositions that have to be assumed. The second assumption, *concept-level consistency*, is based on the fact that even young children are unlikely to make mistakes in tasks involving taxonomic categories (Chi & Ceci, 1987). Thus we assume that, while proofs can have errors, errors in categorical and taxonomic reasoning are less plausible. For example, the consistency constraints that we enforce for proofs prevent propositions such as "velocity of the keys is increasing" and "velocity of the keys is constant" from appearing within the same proof.

A proof is considered sufficiently cheap if the total cost of its assumed atoms is below a certain threshold. The cost is computed for each proposition of the proof via the following procedure. First, costs are uniformly assigned to the goal atoms (observations), namely the propositional representation of the student's essay. Conjunct atoms p_i in the body of a rule have pre-assigned weights

Relation	1st and 2nd arguments	3rd argument	4th argument
`non-equal`	any terms		
`before`	`Time`		
`rel-position`	`Body`	`Rel-location`	
`compare`	`Mag-num` or `D-mag-num` of any scalar or vector quantity	`Ratio`	`Difference`
`compare-dir`	`Dir-num` of any vector quantity	`Rel-dir`	
`dependency`	any terms	`Rel-type`	time interval

Table 2: Relations.

w_i (Stickel, 1988):

$$p_1^{w_1} \wedge \cdots \wedge p_m^{w_m} \rightarrow r_1 \wedge \cdots \wedge r_n.$$

If this rule is used to prove a goal g by unifying it with atom r_j, then the cost of assuming p_i, $1 \leq i \leq m$, is computed according to the following cost propagation formula: $cost(p_i) = cost(g) \cdot w_i$. The cost of the proof is the total cost of all assumed atoms.

A weighted abductive proof for the student's statement "The keys would be pressed against the ceiling of the elevator" is shown in Figure 3. Total cost of the proof is 0.15, the cost of its only assumption. Incidentally, the proof indicates a possible presence of the wrong assumption "The elevator is not in freefall," which is made by the student likely due to a wrong interpretation of a problem given.

Since the cost of a proposition is a penalty for assuming it without a proof, it can also be interpreted as a degree of disbelief in the proposition. This interpretation suggests that more general existentially quantified propositions should be cheaper to assume than more specific propositions. The mechanism for such cost adjustment is implemented in the most recent version of the theorem prover.

Various rule choice heuristics have the aim of finding a sufficiently cheap proof of a small size. Generally, if atoms in the head of the rule are unifiable with a subset of goals then application of such a rule will result in the subset of atoms being removed from the goal list. If a rule has atoms in its body that are unifiable with the goals, then the new subgoals will be *factored* with the unifiable goals, namely only the most specific of the unifiable atoms will be left on the goal list. These nuances imply that proving via rules that have heads and bodies that are unifiable with larger subsets of goals lead to a faster reduction of the goal list and consequently a smaller resultant proof.

In addition, a set of atoms can be cross-referenced via shared variables. The cross-reference graph encodes a large amount of semantics for the proposition corresponding to the respective set of atoms. One of the rule choice heuristics currently being evaluated in the theorem prover is based on the similarity between the graph of cross-references between the propositions in a candidate rule and the graph of cross-references between the set of goals. The metric for the match between two labeled graphs is computed as the size of the largest common subgraph using the decision-tree-based algorithm proposed in (Shearer, Bunke, & Venkatesh, 2001). For further details on the proof search heuristics we refer the reader to (Makatchev, Jordan, & VanLehn, 2004b).

Evaluation

Although students in a baseline evaluation of the Why2-Atlas system showed significant learning gains (VanLehn et al., 2002), the sentence-level representations of the students' essays produced by the system, that are the input to Tacitus-lite+, were too sparse for any misconceptions to be correctly identified. To evaluate Tacitus-lite+ we developed a test suite of 45 student generated essays in which we manually corrected and completed the input generated by the system for input to Tacitus-lite+ and annotated the misconceptions expressed in each essay that Tacitus-lite+ should identify. The student essays were randomly selected from those collected during the baseline evaluation and from subsequent experiments with students and human tutors. In the 45 essays of the test suite, three essays have two misconceptions each, eight essays have one misconception each, and the rest of the essays don't have any misconceptions from the list of 54 misconceptions that could arise for the training problems according to our physics experts.

There are two types of evaluations of interest to us for the abductive theorem prover: (1) the accuracy of the misconceptions revealed by the proofs and (2) the accuracy of the proofs as models of the students. We summarize here the results of both for an earlier version of Tacitus-lite+, as described in (Jordan et al., 2003), and plan to repeat both in the near future for the newer version described in this paper.

To assess the accuracy of the misconceptions identified by the theorem prover, we compare the misconceptions revealed by the proofs of each essay to those annotated for each test suite essay. We accumulated the number of true positives TP, false positives FP, true negatives TN, and false negatives FN for each essay; and from this computed recall TP/(TP+FN), precision TP/(TP+FP), and positive false alarm rate FP/(FP+TN). In addition, we calculated these measures for the theorem prover's results at various proof cost thresholds to see how the performance changes as we move closer toward building a complete proof. The results are shown in Figure 4.

The recall increases from 0 at a proof cost of 1 (where everything is assumed without proof) to 62% at a proof cost threshold of 0.2. As the recall increases, the precision degrades but then levels off. These results mean that the earlier theorem prover can help to reveal up to 62% of the misconceptions that a human would recognize, but at the cost of identifying some misconceptions that are not justified by the essays. We consider recall to be the more important measure for misconceptions since it is important to find and address the misconceptions

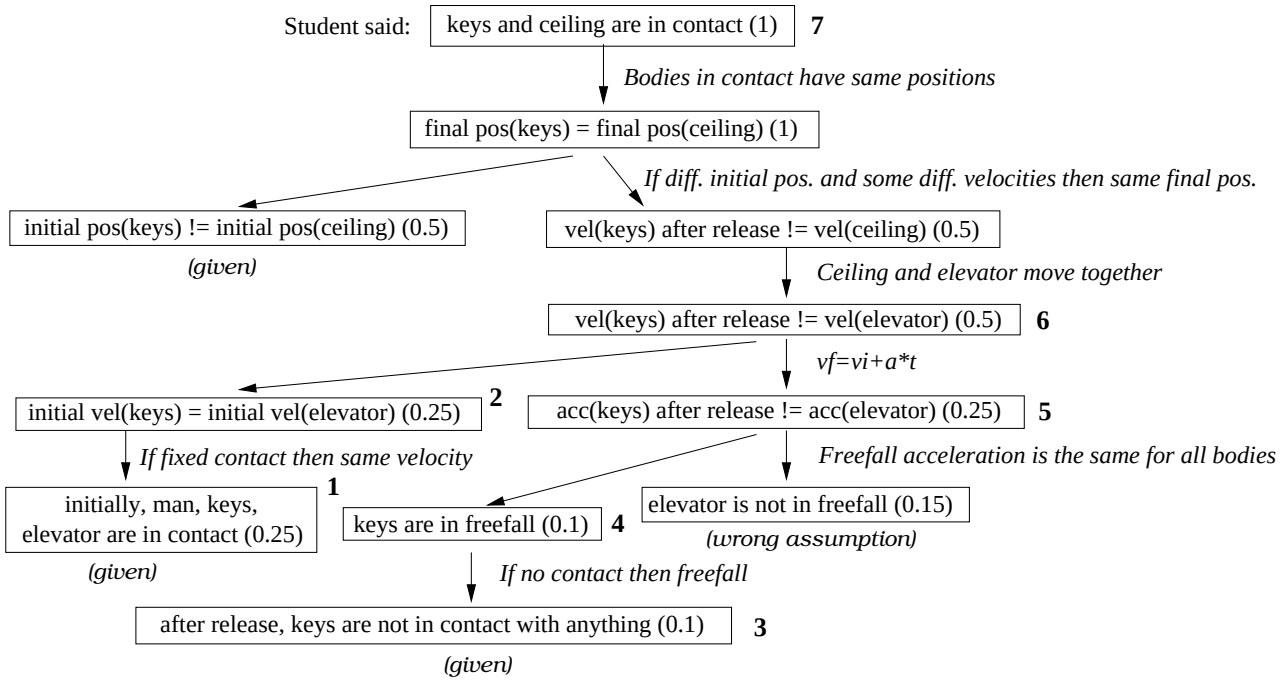

Figure 3: A weighted abductive proof of the proposition representing the excerpt "The keys would be pressed against the ceiling of the elevator." Rule names are in italics; arrows are in the direction of abductive inference; costs of the propositions are in parenthesis; the references to the steps in Figure 2 are in bold. Total cost of the proof is 0.15.

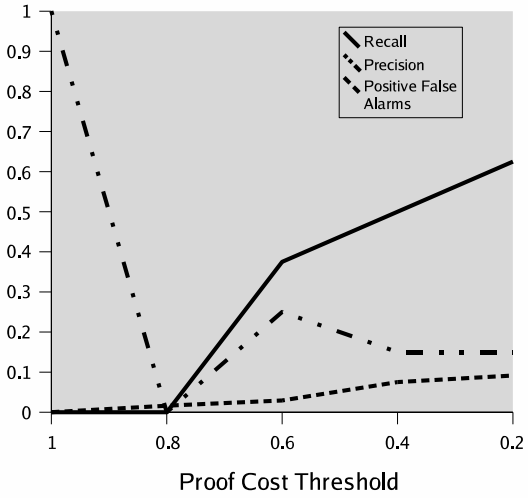

Figure 4: Recall, precision and false alarm measures as proof cost threshold decreases.

that are expected to be obvious to a human tutor. The positive false alarm is quite low and although our goal is to reduce this value as close to 0 as possible, we consider a high recall to be a higher priority as we expect that it is more important not to miss misconceptions. On the other hand, some possible drawbacks of not also trying to lower the positive false alarms are inadvertently strengthening the reasoning that leads to a misconception and a loss of student motivation and cooperation if the student perceives the system is too frequently giving inappropriate feedback.

While these results are encouraging, we expect that the recent improvements we've made to Tacitus-lite+, along with additional testing and fine-tuning of rules, will further improve the results. In addition, an evaluation of misconceptions revealed is only a coarse measure of the quality of the proofs generated. To determine the fitness of the theorem prover's modeling to support assessments of completeness, we must also consider the accuracy of the proof structure generated. Assessing the accuracy of the proof structure is more difficult because the proofs must be hand verified. It is difficult to create a reliable gold standard against which to evaluate the accuracy of proofs for essays and the reasons for any inaccuracy. This is because, in general, language in context gives rise to many inferences (Austin, 1962; Searle, 1975). For this assessment we judged whether the lowest cost proofs generated for 15 of the test suite essays was a plausibly good, satisfactory or bad model of the student essay. As shown in Table 3, as the proof cost threshold decreased and consequently the number of assumptions made fell, the number of good proofs increased and the number of bad ones fell to 0.

Conclusions and Future work

In this paper we described an approach to modeling of a student's reasoning about qualitative physics problems by treating the student's essay as an observation, the problem statement as a set of given facts, and us-

Threshold	0.8	0.6	0.4	0.2
good	7	7	10	11
satisfactory	4	4	4	4
bad	4	4	1	0

Table 3: Evaluation of plausibility of proofs generated for different proof cost thresholds.

ing an abductive proof of this observation as a plausible approximation of the student's reasoning. Abductive proofs provide an intuitively natural representation for logical relations between the arguments of the essay. The problem of insufficient coverage of the domain and common-sense knowledge—one of the difficulties that formal methods face when applied to natural language text analysis—is alleviated by allowing proofs to include assumptions, namely propositions that cannot be proven. Weighted abduction provides a facility to rate such proofs by assigning costs to their respective sets of assumptions. The weighted abductive theorem prover has been implemented and evaluated with respect to plausibility of proofs as models of students' reasoning.

There are a number of challenges still to address. One is to handle various degrees of formalism in the input representations of the student's language. For example, if a student says "throw," the current representation input to Tacitus-lite+ is "apply an upward vertical force." But the student's actual lexical choices need additional reasoning relative to the model of the student in order to determine whether the correct formal representation is plausible for the student. Otherwise, the student is credited with understanding more about physics than may be plausible. So the goal is to take over more of the natural language semantic interpretation process within Tacitus-lite+. The favorable evaluation results we have obtained so far make it more promising that such a move will be successful. Other challenges include increasing the coverage of the rule-base, further improving the efficiency of the theorem prover, and further improved consistency checking.

Acknowledgments

This work was funded by NSF grant 9720359 and ONR grant N00014-00-1-0600. We thank the entire Natural Language Tutoring group, in particular Michael Ringenberg and Roy Wilson for their work on Tacitus-lite+, and Brian 'Moses' Hall and Michael Böttner for their work on knowledge representation and rules.

References

Austin, J. L. (1962). *How to Do Things With Words*. Oxford University Press, Oxford.

Chi, M. T. H., & Ceci, S. J. (1987). Content knowledge: Its role, representation and restructuring in memory development. *Advances in Child Development and Behavior*, *20*, 91–142.

de Kleer, J. (1990). Multiple representations of knowledge in a mechanics problem-solver. In Weld, D. S., & de Kleer, J. (Eds.), *Readings in Qualitative Reasoning about Physical Systems*, pp. 40–45. Morgan Kaufmann, San Mateo, California.

Hobbs, J., Stickel, M., Martin, P., & Edwards, D. (1988). Interpretation as abduction. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pp. 95–103.

Jordan, P., Makatchev, M., & VanLehn, K. (2003). Abductive theorem proving for analyzing student explanations. In *Proceedings of International Conference on Artificial Intelligence in Education*, pp. 73–80, Sydney, Australia. IOS Press.

Kakas, A., Kowalski, R. A., & Toni, F. (1998). The role of abduction in logic programming. In Gabbay, D. M., Hogger, C. J., & Robinson, J. A. (Eds.), *Handbook of logic in Artificial Intelligence and Logic Programming*, Vol. 5, pp. 235–324. Oxford University Press.

Makatchev, M., Jordan, P. W., & VanLehn, K. (2004a). Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems. *To appear in Journal of Automated Reasoning, Special issue on Automated Reasoning and Theorem Proving in Education*.

Makatchev, M., Jordan, P. W., & VanLehn, K. (2004b). Modeling students' reasoning about qualitative physics: Heuristics for abductive proof search. In *Proceedings of Intelligent Tutoring Systems Conference*, LNCS. Springer. To appear.

Ploetzner, R., Fehse, E., Kneser, C., & Spada, H. (1999). Learning to relate qualitative and quantitative problem representations in a model-based setting for collaborative problem solving. *The Journal of the Learning Sciences*, *8*, 177–214.

Reimann, P., & Chi, M. T. H. (1989). Expertise in complex problem solving. In Gilhooly, K. J. (Ed.), *Human and machine problem solving*, pp. 161–192. Plenum Press, New York.

Searle, J. R. (1975). Indirect Speech Acts. In Cole, P., & Morgan, J. (Eds.), *Syntax and Semantics 3. Speech Acts*. Academic Press. Reprinted in *Pragmatics. A Reader*, Steven Davis editor, Oxford University Press, 1991.

Shearer, K., Bunke, H., & Venkatesh, S. (2001). Video indexing and similarity retrieval by largest common subgraph detection using decision trees. *Pattern Recognition*, *34*(5), 1075–1091.

VanLehn, K., Jordan, P., Rosé, C., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., & Srivastava, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of Intelligent Tutoring Systems Conference*, Vol. 2363 of *LNCS*, pp. 158–167. Springer.

Walther, C. (1987). *A many-sorted calculus based on resolution and paramodulation*. Morgan Kaufmann, Los Altos, California.