# Abductive Proofs as Models of Qualitative Reasoning

**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 an application of weighted abductive theorem proving that is used to create a model of students' qualitative reasoning for the Why2-Atlas tutoring system. The system encourages a student to write an essay in natural language so that the essay provides both an explanation as well as an answer to a qualitative mechanics problem. The student's essay is first mapped into a first-order predicate logic representation, which the abductive theorem prover treats as a goal (observation) in order to generate a proof that explains the essay. The resulting proof (1) provides an evaluation of the correctness of the student's essay, and, in the event the essay contains errors, (2) provides a diagnosis of the observed errors that help identify possible tutoring actions. We describe the knowledge representation, rules and weighted abductive theorem proving framework, outline previous and upcoming evaluations, and discuss possible future directions.

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 a verbatim student explanation.

## Introduction

Qualitative physics problems help reveal deep misconceptions in both novices and experienced students (Ploetzner, Fehse, Kneser, & Spada, 1999). The Why2-Atlas tutoring system is designed to encourage students to write their answers to qualitative mechanics problems along with detailed explanations supporting their arguments (VanLehn, Jordan, Rosé, Bhembe, Böttner, Gaydos, Makatchev, Pappuswamy, Ringenberg, Roque, Siler, & Srivastava, 2002). To provide relevant tutoring feedback a deep understanding of student essays is necessary. Consider, for example, the qualitative physics problem shown in Figure 1 along with an actual student explanation. An informal example of a possible chain of reasoning that the student used to arrive at the statement "The keys would be pressed against the ceiling of the elevator" is shown in Figure 2.

In an earlier paper (Jordan, Makatchev, & VanLehn, 2003) we argued that statistical text classification approaches that treat text as an unordered bag of words, e.g. (Landauer, Foltz, & Laham, 1998; McCallum & Nigam, 1998), do not provide a sufficient degree of understanding of the logical structure of a student's essay. Formal approaches to natural language understanding, however, face challenges of their own. Included among the challenges are the need to account for various degrees of formalism and for commonsense knowledge (Dahlgren, McDowell, & Stabler, 1989); and the fact that unconstrained natural language will inevitably exceed the coverage of the knowledge representation (KR) and knowledge base.

Understanding natural language statements in the domain of mechanics has been previously addressed in such physics problem solvers as ISAAC (Novak, 1976) and BEATRIX (Novak & Bulko, 1990). These applications aim at understanding word problems in mechanics and use more constrained language than that of a typical explanation. The task of analyzing students' explanations also differs from the task of problem solving in that there is no need to generate a correct solution. Rather, given a candidate solution, our goals are to evaluate its correctness (validation) and to diagnose possible errors in reasoning (diagnosis). The results from validation and diagnosis determine the tutor's response to the student's essay.

We have chosen to address the goals of validation and diagnosis by treating the student's essay as an observation and generating an abductive diagnosis (Poole, Goebel, & Aleliunas, 1987). A weighted abductive theorem prover (Hobbs, Stickel, Martin, & Edwards, 1988) generates an explanation, a *proof*, possibly making a number of assumptions along the way. The weights for the antecedent atoms of each rule provide a facility for computing the cost of assuming any atom as true without proving it. This enables us to evaluate the cost of each candidate proof at any point during the proof search by computing the total cost of the set of the respective assumptions in the proof. The fact that any atom can be assumed also alleviates the problem of limited knowl-

| 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 | keys' mass is smaller than mass of the elevator | commonsense knowledge |
| 6 | after the release, the keys' acceleration is less than the elevator's acceleration | *A lighter body has a smaller acceleration of freefall |
| 7 | after the release, the keys' velocity is less than the elevator's velocity | if initial downward velocities are the same, then a body with smaller downward acceleration will have smaller downward velocity |
| 8 | 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 rule is preceded by an asterisk.

edge base coverage, by allowing for evaluation of proofs of observations even in cases when the knowledge base is insufficient to explain all observed atoms. The cheapest proof found within thresholds on cost and time is chosen as a plausible model of the student's reasoning that explains the statements of the essay. The validation task is then reduced to checking if the proof contains any false assumptions. The diagnosis is the set of the false assumptions in the proof. This approach resembles the diagnostic application proposed for model builders like MISQ (Richards, Kraan, & Kuipers, 1992), but differs in the fact that the input in our application includes not only the behavior of a system but also an explanation of the behavior.

The theorem prover we use, called Tacitus-lite+, is a derivative of SRI's Tacitus-lite (Hobbs et al., 1988, p. 102) that incorporates a number of extensions, including sorts and proof search heuristics. To account for different degrees of formalism that students use in solutions and explanations of textbook qualitative mechanics problems, the rule base has to include both qualitative physics rules in the domain of mechanics as well as rules for translating commonsense knowledge into proper physics representations (defined as *idealization rules* in the next section). To support the task of validation, qualitative rules should be precise enough to ensure soundness of the theory (so that erroneous essays do not get good abductive explanations). In the section on rules we discuss these issues in more detail.

In the following section we address the knowledge representation and rules for reasoning in the domain of qualitative mechanics. We then present an overview of the abductive theorem proving framework and proof search heuristics. 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 et al., 2003). Finally, we conclude with a summary of the results and an outline of future directions.

## 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 et al., 1999; Reimann & Chi, 1989). In (Reimann & Chi, 1989) the internal representation is described in terms of "objects, operators, and constraints, as well as initial and final states." This notion of internal representation overlaps with the notion of a *path* in an *envisionment* (de Kleer, 1990), i.e. a particular behavior from the set of all possible behaviors of the system. We refer to a further step, that translates a path in the envisionment into the domain terminology (bodies, forces, motion properties), as an *idealization* (Makatchev, Jordan, & VanLehn, 2004a).

For the problem in Figure 1, for example, a possible path in the envisionment is as follows: (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 of this path would be:

Bodies: Keys, Man, Elevator, Earth.

Forces: Gravity, Man holding keys

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

Many misconceptions that students have are rooted in envisionment and idealization (Ploetzner et al., 1999). To make the task of representing possible correct and erroneous paths in envisionments feasible, we restrict ourselves to problems where envisionments have few plausible paths. The correct and buggy rules of mechanics (which rely mostly on the formal domain terminology) are augmented by correct and buggy rules for reasoning about idealization (which translate loose language into formal terms). Further we describe our approach to representing the statements and rules. A more complete coverage of this material can be found in (Makatchev et al., 2004a).

## Qualitative mechanics ontology

The ontology for the subset of qualitative mechanics that the system addresses builds upon (Ploetzner & VanLehn, 1997) and consists of bodies (e. g., keys, man), agents (air), phenomena (e. g., gravity, friction), and physical quantities (e. g., force, velocity, position). To adequately represent justifications, we also have representations for references to physics laws (Newton's First Law) and to 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 idealization rules are described in the following section.

Statements about the student's beliefs about a physical model are represented in a first-order predicate language. Logical constants and variables, that correspond to bodies, agents, and quantities, are "typed," that is, associated with a sort. Sorts are partially ordered by a natural subset order. Domains of the arguments of the predicate symbols are restricted to certain sorts. These associations and constraints constitute an *order-sorted signature* (Walther, 1987).

Time is represented using time instants as basic primitives. Time intervals are denoted as a pair $(t_i, t_j)$ of instants. This representation, together with the relation `before` on time instants, allows us to implement the semantics of open time intervals. We do not currently implement the semantics of limits corresponding to "just before" and "just after."

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. Thus, the statement "The keys' vertical acceleration is constant and negative" is represented as a single atom augmented with its order-sorted signature as follows:

```
((acceleration a1 keys vertical constant ?d-mag-num
        ?mag-z ?mag-num neg ?dir-num ?d-dir ?t1 ?t2)
(Quantity1b Id Body Axial Constant D-mag-num
      Mag-zero Mag-num Dir Dir-num D-dir Time Time))
```

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 of `detached`, `attached`, and `moving-contact` (for the case of relative motion between bodies in contact).

The atoms can be cross-referenced via shared variables, as in the representation of the equality of positions shown in Figure 3. Another example of cross-referencing is the representation of the statement "Force of gravity acting on the keys is constant and nonzero":

| Description | Sort |
|---|---|
| quantity | `Quantity1b` |
| identifier | `Id` |
| body | `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`.

```
((force f1 ?body1 keys ?comp constant ?d-mag-num
      nonzero ?mag-num ?dir ?dir-num ?d-dir ?t1 ?t2)
(Quantity2b Id Body Body Comp Constant D-mag-num
      Mag-zero Mag-num Dir Dir-num D-dir Time Time))
((due-to d1 f1 ph1)  (Due-to Id Id Id))
((phenomenon ph1 gravity)
(Phenomenon Id Field-interaction))
```

In the example above, the predicate `due-to` is used to refer to the phenomenon responsible for the force. Roles of the forces, such as centripetal, reaction and weight are represented using the predicate `role`.

To account for the fact that the qualitative problems can often be solved with or without an explicit definition of a coordinate system, the direction of vector quantities can be specified in two ways: (1) in terms of a fixed coordinate system of the problem, i.e. "vertical velocity is negative", and (2) in a loose language, e.g. "velocity is down." Similarly, "keys are moving down" is represented as a `motion` atom, equivalent to "keys' motion is downward", and an idealization rule translates it into a more formal `velocity` atom, namely "vertical velocity is negative," which introduces assumptions about the student's understanding of the respective inferences.

## Rules

As we mentioned above, we would like the system to reason about an idealized model as well as about the process of idealization. In fact, the ideal solutions produced by our physics experts are broken into two stages: (1) defining relevant bodies, motion, and forces in physics terms (a model), and (2) applying physics principles within the model to derive the answer. To support the task of reasoning about these stages, we have three classes of rules: givens, idealization rules, and rules of qualitative mechanics. The correct idealization of the problem statement is represented as a set of givens for the theorem prover, namely as rules of the form $\rightarrow a$. An idealization atom $b$ that allows for the buggy counterpart $b'$ is represented by a pair of the rules $\rightarrow b$ and $\mathit{bug\_b} \rightarrow b'$, where atom $\mathit{bug\_b}$ is not in the head of any rule (i.e. it has to be assumed).

Similarly rules that have buggy counterparts are represented by the pair $a \rightarrow b$ and $\mathit{bug\_ab} \wedge a' \rightarrow b'$. Note that since atoms in the head of a rule can include vari-

| 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.

ables, the rule or the given can be used to prove different goals. Since, even within the same essay, a rule can be used both correctly and incorrectly, we do not declare all pairs of such rules as mutually exclusive. Instead we enforce the exclusiveness of rules selectively at the meta-level of the prover. When the application of a rule would generate a new goal atom that is inconsistent with an atom that has already been proven (due to functional properties of predicate arguments), it is excluded due to the consistency constraints (see the following section).

Examples of idealization rules have been discussed in the previous section. This class of rules covers such inferences as "Distance between two bodies is decreasing → the bodies are said to be 'closer'," "A body's vertical velocity is positive and vertical axis is directed upwards → the body's velocity is directed upward."

Mechanics rules cover correct and buggy reasoning at the level of an idealized model of the problem, for example, "Zero acceleration → constant velocity," and "Zero force → decreasing velocity". Many physics problems that are qualitative in nature require reasoning about quantitative proportionalities, e.g. "Twice as much total force on a body →twice as much acceleration." Although our KR allows us to represent such rules, currently we limit our problems to those that require only qualitative proportionalities, e.g. "More total force on a body → more acceleration."

The rules are represented as *extended Horn clauses*, namely the head of the rule is an atom or a conjunction of multiple atoms. An example of a correct rule, stating that "if the velocity of a body is zero over a time interval then its initial position is equal to its final position", is shown in Figure 3.

As we mentioned in the Introduction, we would like to ensure that the set of correct solutions and explanations of the mechanics problems is covered by a sound rule base, i.e. while a correct essay should be provable by at least one correct abductive proof, it should be impossible for an erroneous essay to be diagnosed as correct unless assumptions are made. In some domains, it appears that only using sound qualitative rules is not sufficient to cover the commonsense conclusions that arise (Forbus, 1997). In the case of the textbook mechanics problems that we have chosen to address, our experts' solutions share a common feature: Once the idealization is performed, most of the inferences are carried out within the realm of the idealized model by sound qualitative versions of physics principles. Therefore, at least as far as the coverage of the reasoning within the idealized model is concerned, we can limit ourselves to using sound rules.

At the time of the evaluation presented in this paper (Summer 2003), the knowledge base consisted of 24 idealization rules (excluding problem-specific givens that are assumed to be shared knowledge), 24 buggy rules, and 57 rules of qualitative Newtonian mechanics.

```
((velocity v1 ?body ?comp ?d-mag ?d-mag-num
          0 ?mag-num ?dir ?dir-num ?d-dir ?t1 ?t2)
(Quantity1b Id Body Comp D-mag D-mag-num
      Mag-zero Mag-num Dir Dir-num D-dir Time Time))
→
((position p1 ?body ?comp ?d-mag1 ?d-mag-num1
  ?mag-z1 ?mag-num1 ?dir1 ?dir-num1 ?d-dir1 ?t1 ?t1)
(Quantity1b Id Body Comp D-mag D-mag-num
      Mag-zero Mag-num Dir Dir-num D-dir Time Time))
((position p2 ?body ?comp ?d-mag1 ?d-mag-num1
  ?mag-z1 ?mag-num1 ?dir1 ?dir-num1 ?d-dir1 ?t2 ?t2)
(Quantity1b Id Body Comp D-mag D-mag-num
      Mag-zero Mag-num Dir Dir-num D-dir Time Time))
```

Figure 3: Representation for the rule "If the velocity of a body is zero over a time interval then its initial position is equal to its final position." Atoms are augmented with their respective sorted signatures.

## 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 assumptions) and $I$ is a set of integrity constraints. Then an *abductive explanation* of a given set of sentences $G$ (goals) is (a) a subset $\Delta$ 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 $\Delta$ 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 $\tau_i$. 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:

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

### Proof search heuristics

The aim of the proof search heuristics is to quickly find a proof that optimizes a combination of 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 could 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 $w_i$ (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 4. Total cost of the proof is 0.22, the cost of its two assumptions. Incidentally, the proof indicates a possible application of the buggy rule "A lighter body has a smaller acceleration of freefall," which is a common misconception.

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 achieving those goals. On the other hand, if a rule has atoms in its body that are unifiable with the goals, then the new subgoals from the body will be *factored* (combined via unification) 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 in a rule or in the goal list can be cross-referenced via shared variables (see the section on qualitative mechanics ontology). 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 atoms in a candidate rule and the graph of cross-references between the atoms on the goal list. 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 first-order predicate logic representation 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 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 pilot studies with 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 have none of the 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

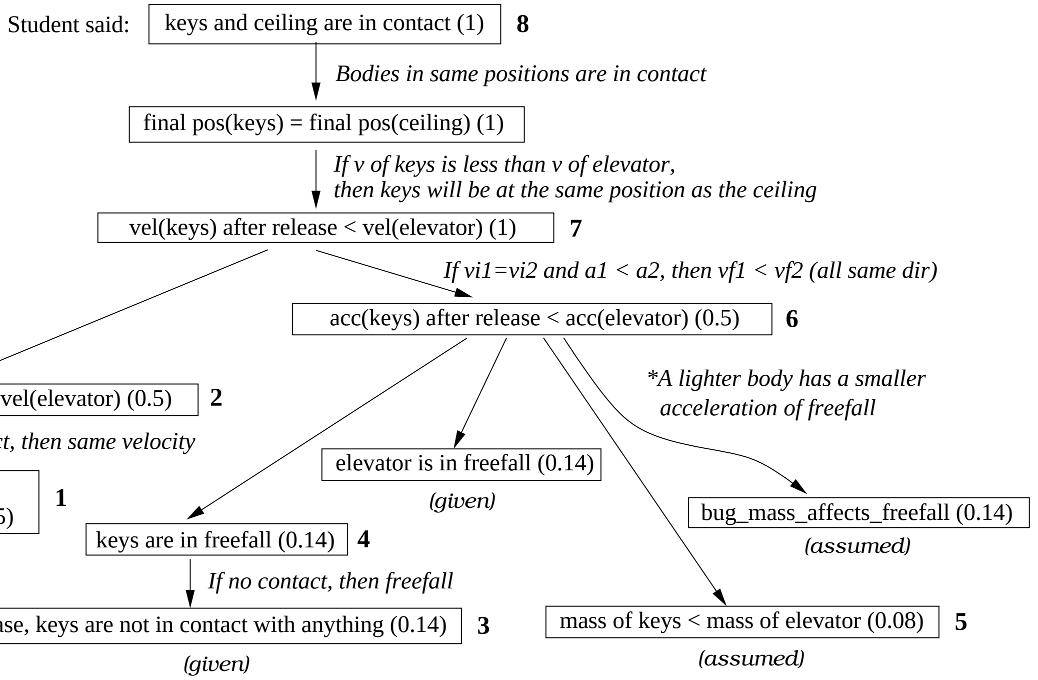
Student said:
keys and ceiling are in contact (1)
8
Bodies in same positions are in contact
final pos(keys) = final pos(ceiling) (1)
If v of keys is less than v of elevator,
then keys will be at the same position as the ceiling
vel(keys) after release < vel(elevator) (1)
7
If vi1=vi2 and a1 < a2, then vf1 < vf2 (all same dir)
acc(keys) after release < acc(elevator) (0.5)
6
*A lighter body has a smaller
acceleration of freefall
initial vel(keys) = initial vel(elevator) (0.5)
2
If fixed contact, then same velocity
elevator is in freefall (0.14)
(given)
initially, man, keys,
elevator are in contact (0.5)
1
(given)
bug_mass_affects_freefall (0.14)
(assumed)
keys are in freefall (0.14)
4
If no contact, then freefall
after release, keys are not in contact with anything (0.14)
3
(given)
mass of keys < mass of elevator (0.08)
5
(assumed)

Figure 4: 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; the buggy rule is preceded by an asterisk; 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 0.22.

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 5.

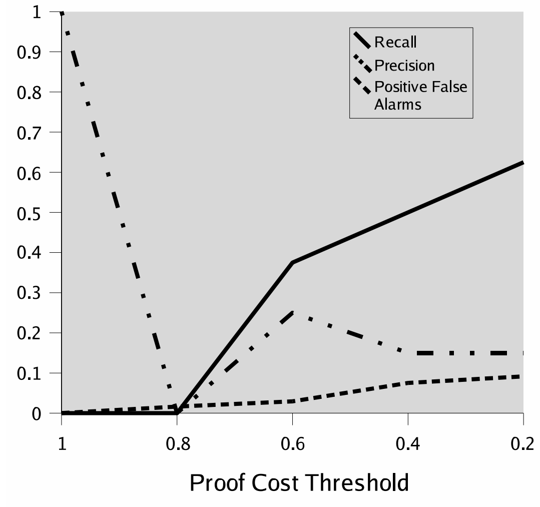

Figure 5: Recall, precision and false alarm measures as proof cost threshold decreases.

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 that are expected to be obvious to a human tutor. The positive false alarms are 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 misconcep-

| 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 the accuracy of the proof structures generated for different proof cost thresholds.

tion 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 with misconceptions is only a coarse measure of the quality of the proofs generated. A more refined measure of the plausibility of proofs as models of the students should take into account 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 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 using an abductive proof of this observation as a plausible approximation of the student's reasoning. Abductive proofs provide an intuitively natural representation for the logical relations between the arguments of the essay. The problems of insufficient coverage of the domain and of common-sense knowledge—two difficulties that formal methods face when applied to natural language text analysis—were 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. This facility can also be viewed as a soft closed-world constraint: cheaper proofs are generally preferred. The challenge of mixed usage of formal physics terminology and loose language in natural language explanations was addressed via idealization rules that translate representations of the latter into representations of the former. Finally we described the adaptations we made to the weighted abductive theorem prover and evaluated the plausibility of the proofs it generated as models of students' reasoning.

The challenges are, however, far from having been conquered. Consider the following situation: If a student says "throw," the current representation that is input to Tacitus-lite+ is "apply a 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. This implies that more natural language semantics interpretation should be postponed and done within the context of the student model. However, doing such processing via idealization rules raises the problems of reasoning at this level that we tried to avoid in our system: (1) explicit representation of large amounts of commonsense knowledge, and (2) the difficulty of providing a set of sound qualitative rules that cover commonsense conclusions (Forbus, 1997).

## 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.

Dahlgren, K., McDowell, J., & Stabler, Jr., E. P. (1989). Knowledge representation for commonsense reasoning with text. *Computational Linguistics*, *15*(3).

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.

Forbus, K. D. (1997). Qualitative reasoning. In Tucker, Jr., A. B. (Ed.), *The Computer Science and Engineering Handbook*. CRC Press.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

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.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceeding of AAAI/ICML-98 Workshop on Learning for Text Categorization*. AAAI Press.

Novak, G. S. (1976). Computer understanding of physics problems stated in natural language. *American Journal of Computational Linguistics*.

Novak, G. S., & Bulko, W. (1990). Understanding natural language with diagrams. In *Proceedings of the AAAI-90*, pp. 465–470.

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.

Ploetzner, R., & VanLehn, K. (1997). The acquisition of qualitative physics knowledge during textbook-based physics training. *Cognition and Instruction*, *15*(2), 169–205.

Poole, D. L., Goebel, R., & Aleliunas, R. (1987). Theorist: a logical reasoning system for defaults and diagnosis. In McCalla, G. I., & Cercone, N. (Eds.), *The Knowledge Frontier: Essays in the Representation of Knowledge*, pp. 331–352. Springer, New York.

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.

Richards, B. L., Kraan, I., & Kuipers, B. J. (1992). Automatic abduction of qualitative models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-92)*. AAAI/MIT Press.

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.