Design and Evaluation of a Natural Language Tutor for Force and Motion

Ryan Badeau and Andrew F. Heckler

Department of Physics, The Ohio State University, 191 W. Woodruff Avenue, Columbus, OH 43210

Abstract: We report on the design and pilot evaluation of a simple natural language computer tutor that targets student difficulties with the concepts of force and motion. The tutor prompts students to respond in free-response natural language to questions that address the relationships between the directions of net force, velocity, and acceleration. To examine the effectiveness of the natural language format, we compared student performance on a previously validated force and motion assessment after tutoring via natural language and multiple choice formats. Natural language training with feedback, multiple choice training with feedback, and natural language training without feedback formats resulted in effect sizes of $d = 0.60$ ($p = 0.07$), $d = 0.46$ ($p = 0.13$), and $d = 0.09$ ($p = 0.97$) respectively versus a no-training control. In addition, a median split on course grades showed no significant aptitude-treatment interaction across training conditions. However, accounting for time spent on training, the multiple choice training was significantly more efficient. For the natural language format, an analysis of performance (62% identification of an initial student response), false positives, and typical student answer patterns suggest room for improvement and subsequent study.

INTRODUCTION

The motivation behind the development of conversational computer tutors can be traced to the success of one-on-one tutoring methods. Physics education, in particular, has an impressive pedigree of development – including ANDES/ATLAS [1], the AutoTutor series [2,3], Cordillera [4], and most recently Deep Tutor [5]. These computer tutors, using various natural language methodologies and to significant levels of success, have tackled physics topics such as forces, kinematics, Newton's laws, and energy conservation. The approaches vary from guided essay construction focused on the identification of misconceptions [2,3], to specific knowledge construction dialogues [1,4]. The common hope is that by engaging the student in a reflective and constructive dialog, students will achieve greater learning gains than by rote application of physical principals.

In light of these successes, we chose to target a foundational subset of Newtonian Mechanics – specifically, a novel, systematic focus on the relationships between the directions of net force, velocity, and acceleration in one dimension. The motivation for focusing in on these specific conceptual relations is two-fold. First, recent work has suggested the possibility for empirically-validated learning progressions between these relationships [6], a finding that evokes potential applications for an intelligent tutoring system. Second, the concepts of force and motion in physics are among the first introduced to novices, and come with clear conceptual and conversational baggage. Conceptual issues such as the assumption of a force in the direction of motion are both pervasive and persistent [7]. Moreover, these conceptual issues are potentially compounded by the use of imprecise every-day language transferred to the physics setting – a student may refer to an object "moving", and it is not immediately clear to the listener (and perhaps even the student), whether they are implying something about the object's velocity, acceleration, or both. On the other hand, the precise use of language – or the ability to "talk like a physicist" is often heralded as one of many content-external goals of physics education.

Consequently, our ultimate goal is to develop a system that is adaptive, aligns with empirically-validated learning progressions, and utilizes the strengths of natural language dialog – where most applicable – to help students efficiently master these concepts. As a first step towards this goal, we piloted a simple natural language tutor, in order to compare the performance of natural language to a traditional multiple choice format during training on the relationships between the directions of net force, acceleration, and velocity.

As the intention of this investigation is to explore the combination of the free-response natural language question format with immediate feedback in order to elicit and confront specific student difficulties, and not to pioneer novel natural language processing techniques, the natural language implementation used

here is based in part on the established statistical technique of latent semantic analysis (LSA). LSA is a high-dimensional, corpus-based method of analyzing similarity between any two pieces of text – words, phrases, sentences, or entire documents [8]. In addition to being utilized in the AutoTutor series, LSA has been implemented in multiple other intelligent tutoring systems [2]. In particular, this initial study utilizes the libraries and functionality provided by Gensim, an open source python semantics package [9].

TUTOR DESIGN

The tutor interface, shown in Fig. 1, is the same for both the natural language and multiple choice formats. The interface consisted of a scrolling-dialog window (left) where the question, student response, and feedback were shown; a window for question related graphics (upper right); a session progress-bar (right); a student input line (bottom left); and an optional upvote/downvote prompt to measure student opinion on the usefulness of provided feedback (bottom right).
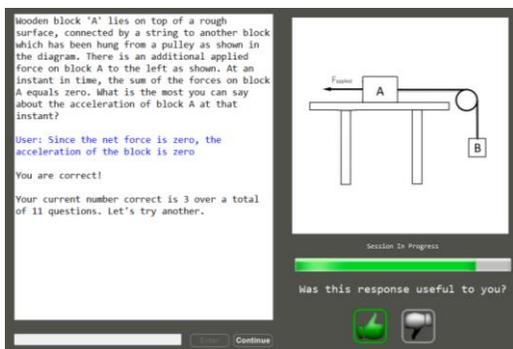

FIGURE 1. The tutorial interface used for all formats.

The structure of a question-turn in the tutorial was as follows: a question was posed, the student responded either in multiple choice or free-response natural language format, and the tutorial provided corresponding feedback. In the event of a correct response, the tutorial simply remarked "You are correct!" and prompted the student to continue. In the event of an incorrect response, the tutorial provided the correct answer, along with a one or two sentence explanation. The feedback was identical between the two conditions. An example question, student response and feedback turn in each format are shown in Table 1.

METHOD AND PARTICIPANTS

A total of 171 participants were randomly assigned to one of four training conditions: natural language format (N=40), multiple choice format (N=45), natural

language format without feedback (N=48), and a no-training control (N=38). The participants were students enrolled in the second semester of a calculus-based introductory physics sequence at The Ohio State University, a large public research university. Course credit was given for participation. A one-way ANOVA showed no statistical differences between students' overall course grades across the four conditions, $F(3,167) = 0.087$, $p = 0.97$.

The training consisted of 12 questions targeting the relationships between the directions of net force, velocity, and acceleration. Participants completed one of the training conditions, followed immediately by a multiple-choice assessment (with no feedback). The 16 question assessment was composed of questions from the diagnostic developed previously by Rosenblatt and Heckler [6]. Participants completed the training and assessment in individual carrels in a quiet testing room.

TABLE 1. Example question, student response, and provided feedback from both question formats.

Natural Language	Multiple Choice
One of Ohio State's football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?	One of Ohio State's football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?
User: the velocity is towards the offensive line	(a) His velocity is directed towards the offensive line.
	(b) His velocity is directed away from the offensive line.
Sorry, but your answer is not correct.	(c) His velocity is zero.
	(d) His velocity could be directed either towards or away from the offensive line.
The correct answer is that the velocity could be directed towards the offensive line, away from the offensive line, or it could be zero. The acceleration does not require that the velocity have any particular direction at that instant. For instance, the player could be at rest and accelerating towards the line, or accelerating towards the line but with an instantaneous velocity directed away (that is moving away and slowing down).	(e) His velocity could be directed towards the offensive line, or it could be zero.
	(f) His velocity could be directed towards the offensive line, away from the offensive line, or it could be zero.
	User: a
	Sorry, but your answer is not correct.
	<…feedback continued…>

RESULTS

Overall Performance

The mean-scores on the assessment are shown in Fig. 1 for each of the four training conditions: natural language format (68%), multiple choice format (66%), natural language format without feedback (57%), and the no-training control (54%). A one-way ANOVA indicated a significant difference between the four conditions ($F(3,167) = 3.2$, $p = 0.03$).

A Tukey post-hoc test shows a marginally significant difference in scores between the natural language format and control ($p = 0.07$, $d = 0.60$), but not between multiple choice format and control ($p = 0.13$, $d = 0.46$), nor between natural language without feedback and control ($p = 0.97$, $d = 0.09$). There was no significant difference in scores between the natural language and multiple choice formats.
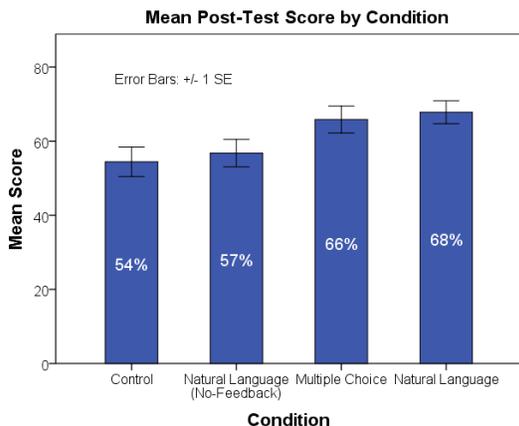

FIGURE 2. Mean score for each training condition.

To check for an aptitude-treatment interaction, we divided students using a median split on their final course grades. Using this split, a 2 (upper vs. lower course grade) x 4 (training condition) ANOVA showed main effects from course grade ($p < 0.001$), and condition ($p = 0.02$), but no significant interaction effect. In short, students who performed better overall in the second semester physics course consistently performed better on the assessment, regardless of training condition. This suggests that the simple natural language format piloted here neither preferentially helps those who perform better in the course (who may already have a better grasp of the expected answer and language), nor those who perform worse (who may be the most to benefit) compared to stand-alone instruction or multiple choice training.

Efficiency

The median times spent on the training were 804.5s for the natural language format, 682.5s for the natural language format without feedback, and 487.5s for multiple choice format. A median test showed that the training time was significantly different between conditions ($\chi^2 = 18$, $p < 0.001$).

In order to better compare potential trade-offs between learning gains and training time, we define the efficiency of training for a particular student as the ratio of assessment score over total time spent during training (in minutes). The mean efficiency ratings (score/min) for each of the training methods were 8.44

for multiple choice format, 5.16 for natural language, and 5.52 for natural language without feedback. A one-way ANOVA showed that the difference between these efficiencies is significant ($F(2,130) = 11.3$, $p < 0.001$). A Tukey post-hoc showed that the multiple choice format was significantly more efficient than either the natural language or natural language without feedback condition ($ps < 0.001$). The difference between efficiency for natural language and natural language without feedback was not significant ($p = 0.89$).
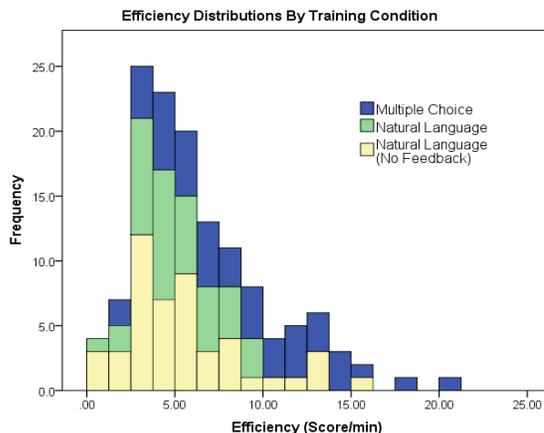

FIGURE 3. Efficiency distributions for each training condition.

The distribution of individual efficiency ratings for each condition is shown in Fig. 3. The main finding to note is that a large part of the success of the multiple choice format – at least as defined by this efficiency metric – is the long tail of students who performed well on the assessment and proceeded quickly through the corresponding training. A similar tail for the natural language without feedback condition helps to explain the comparable efficiency performance with the natural language format. In essence, although the natural language format with feedback resulted in comparable overall performance, these gains resulted from about a 40% time trade-off (here, approximately 5 minutes).

Language Accuracy and Student Answer Patterns

One of the goals of this pilot study was to analyze the accuracy and effectiveness of our simple natural language implementation and identify potential improvements. To begin, it is worth noting that out of 40 students in the natural language condition, 2 students demonstrated identifiable gaming behavior. In both cases, the student rapidly (< 10 seconds spent on the natural language format question) entered a short noncommittal phrase like "inconclusive" or submitted

a blank natural language response on more than one training question. Discarding the responses of these two students, the natural language model produced a match for 62% of student responses – meaning 38% of the time, the natural language tutor was unable to successfully match the initial student answer to an answer prototype. In such cases, the student was asked a clarifying multiple choice version of that particular question. In addition to limitations of the LSA technique and simple cases of vague wording, typical student answer patterns suggest several additional failure-points in identifying student responses.

First, students would frequently invoke ambiguous coordinate systems in their free-responses to describe the direction of a physical quantity of interest. For example, a response would refer to a "negative" direction, or a particular vector being "positive", even though no coordinate axis was specified in the problem or explicitly stated by the student.

Second, students would occasionally state partially correct relationships or definitions, which although relevant to the question, were not necessarily what the question was asking. For example, a student would correctly define Newton's Second Law, but then not apply it to find the direction of the acceleration. Another student, having correctly stated that acceleration was the time rate of change of velocity, incorrectly determined the resultant direction of the object's velocity. Such cases suggest the potential for targeted clarifying dialog, perhaps similar to the knowledge-construction-dialogs used elsewhere [1,4].

For those statements which the natural language tutor produced a match, we tracked the number of false-positives and false-negatives – instances where the tutor incorrectly declared to the student that a response was correct, or incorrectly declared a response was not correct respectively – by comparing the computer identification to hand-coded grading. We find an overall false-positive rate of 3.5%, largely driven by one question with a false-positive rate of 7.9%, and a false-negative rate of 9.3%. To test for any negative learning effects of false-positives (false-negatives were less of a concern because of the availability of explanatory feedback), we found that the mean assessment scores between those students who had seen at least one false-positive was not significantly different from those that saw no such misidentifications (71.5% vs. 65.3% respectively, $t(36) = -0.83$, $p = 0.41$).

CONCLUSION

In this pilot study we implemented a simple natural language tutor and demonstrated that it is at least as effective as simple multiple choice practice for learning basic force and motion concepts. However, given the similarity in effectiveness for the natural language and multiple choice conditions with feedback ($p = 0.07$, $d = 0.60$ and $p = 0.13$, $d = 0.46$ respectively) and the minimal impact of eliciting natural language student responses in the absence of feedback ($p = 0.97$, $d = 0.09$), these findings primarily support the value of immediate and specific feedback, rather than recommend a particular question format.

Moreover, although the natural language format elicited useful student responses, the overall performance of the natural language implementation in analyzing those responses remains a potential area for improvement, especially in regards to limiting any potential negative effects from misidentifications on student learning and affect. Whereas this investigation sought to compare the natural language and multiple choice formats in equivalent a form as possible, it may be that natural language necessitates different questions and multi-step refinement of a student response in order to be relatively more effective than a multiple choice counterpart. In fact, the student responses discussed here suggest some simple examples where natural language could help clarify an incorrect response.

Finally, answer selection via the multiple choice format was significantly more time efficient than eliciting a natural language statement from the student. Therefore, given that practice via a natural language tutor is likely to take significantly more time than multiple choice practice, it seems any future gains expected of more refined natural language must target difficulties where multiple choice is insufficient and/or have relatively more affective benefits in order to be worthwhile.

REFERENCES

1. C.P. Rosé et al., *Proceedings of AI in Education 2001 Conference*, 151-153 (2001).
2. A.C. Graesser et al., *Behavior Research Methods, Instruments, & Computers* **36**, 180-192 (2004)
3. A.C. Graesser et al., *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, 1-6 (2003)
4. M. Chi, K. VanLehn, D. Litman, and P. Jordan, *International Journal of Artificial Intelligence in Education* **21**, 83-113 (2011)
5. V. Rus, S K. D'Mello, X. Hu, and A.C. Graesser, *AI magazine* **34**, 42-54 (2013).
6. R. Rosenblatt and A.F. Heckler, *Phys. Rev. Special Topics – Physics Education Research* **7**, 020112 (2011).
7. J. Clement, *American Journal of Physics* **50**, 66 (1982).
8. T.K. Landauer, P.W. Foltz, and D. Laham, *Discourse Processes* **25**, 259-284 (1998)
9. R. Řehůřek and P. Sojka, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50 (2010)