# Annotating Students' Understanding of Science Concepts

**Rodney D. Nielsen, Wayne Ward, James H. Martin and Martha Palmer**
Center for Computational Language and Education Research
Institute of Cognitive Science
Department of Computer Science
University of Colorado, Boulder
Rodney.Nielsen, Wayne.Ward, James.Martin, Martha.Palmer@Colorado.edu

**Abstract**

This paper summarizes the annotation of fine-grained entailment relationships in the context of student answers to science assessment questions. We annotated a corpus of 15,357 answer pairs with 145,911 fine-grained entailment relationships. We provide the rationale for such fine-grained analysis and discuss its perceived benefits to an Intelligent Tutoring System. The corpus also has potential applications in other areas, such as question answering and multi-document summarization. Annotators achieved 86.2% inter-annotator agreement (Kappa=0.728, corresponding to substantial agreement) annotating the fine-grained facets of reference answers with regard to understanding expressed in student answers and labeling from one of five possible detailed relationship categories. The corpus described in this paper, which is the only one providing such detailed entailment annotations, is available as a public resource for the research community. The corpus is expected to enable application development, not only for intelligent tutoring systems, but also for general textual entailment applications, that is currently not practical.

## 1. Introduction

Determining whether the propositions in one text fragment are entailed by those in another fragment is important to numerous NLP applications. Consider an intelligent tutoring system (ITS), where it is critical for the tutor to assess which specific facets of the desired or reference answer are entailed by the student's answer. Truly effective interaction and pedagogy is only possible if the automated tutor can assess this entailment at a relatively fine level of detail (c.f. Jordan et al., 2004).

Still, most ITSs today provide only a shallow assessment of the learner's comprehension (e.g., a correct versus incorrect decision). Many ITS researchers are striving to provide more refined learner feedback (Graesser et al., 2001; Jordan et al., 2004; Peters et al., 2004; Roll et al., 2005; Rosé et al., 2003; VanLehn et al., 2005); however, they are developing very domain-dependent approaches, requiring a significant investment in hand-crafted logic representations, parsers, knowledge-based ontologies, and or dialog control mechanisms. Similarly, research in the area of scoring constructed responses to short answer questions (e.g., Callear et al., 2001; Leacock, 2004; Mitchell et al., 2003; Pulman and Sukkarieh, 2005) also relies heavily on hand-crafted pattern rules, rather than being designed with the goal of accommodating dynamically generated, previously unseen questions and does not provide feedback regarding the specific aspects of answers that are correct or incorrect.

The PASCAL Recognizing Textual Entailment (RTE) challenge (Dagan et al., 2005) is addressing the task of domain independent inference, but the task only requires systems to make yes-no judgments as to whether a human reading one text snippet would typically consider a second text to most likely be true in its entirety. This paper discusses some of the extensions necessary to the RTE scheme in order to satisfy the requirements of an ITS, provides a report on our efforts to produce such an annotated corpus, and presents results of an initial automated classifier.

## 2. The Necessity of Finer-grained Analysis

In order to optimize learning gains in the tutoring environment, there are myriad issues the tutor must understand regarding the semantics of the student's response. Here, we focus strictly on drawing inferences regarding the student's understanding of the low-level concepts and relationships or facets of the reference answer. We use the word facet throughout this paper to generically refer to some part of a text's meaning, most commonly the meaning associated with a syntactic dependency.

Imagine that you are an elementary school science tutor and that rather than having access to the student's full response to your questions, you are simply given the information that their answer was correct or incorrect, a yes or no entailment decision. Assuming the student's answer was not correct, what question do you ask next? What follow up question or action is most likely to lead to better understanding on the part of the child? Clearly, this is a far from ideal scenario, but it is roughly the situation within which many Intelligent Tutoring Systems exist today.

Rather than have a single yes or no entailment decision for the reference answer as a whole, (i.e., does the student understand the reference answer in its entirety or is there some unspecified part of it that we are unsure whether the student understands), we break the reference answer down into what we consider to be its lowest level compositional facets. This roughly translates to the set of triples composed of labeled (typed) dependencies in a dependency parse.[1] The following illustrates how a sim-

[1] In a dependency parse, the syntactic structure of a sentence is represented as a set of lexical items connected by binary directed modifier relations called dependencies. The goal of most English dependency parsers is to produce a single projective tree structure for each sentence, where each node represents a word in the sentence, each link represents a functional category relation, often labeled, between a governor (head) and a subordinate (modifier), and each node has a single governor (c.f., Nivre and Kubler, 2006). Each dependency can be labeled with a type, (e.g., subject, object, nmod –noun modifier, vmod–verb modifier, sbar–subordinate or relative clause).

ple reference answer (1) is decomposed into the answer facets (1a-d) derived from its dependency parse (see Figure 1), with (1a'-d') providing a gloss of each facet's meaning. As can be seen in 1b and 1c, the dependencies are augmented with thematic roles (e.g., Agent, Theme, Cause, etc.; c.f., Kipper et al., 2000). The facets also include those semantic role relations that are not derivable from a typical dependency parse tree. For example, in the sentence "As it freezes the water will expand and crack the glass", *water* is not a modifier of *crack* in a typical dependency tree, but it does play the role of Agent in a semantic parse.

(1) A long string produces a low pitch
(1a) NMod(string, long)
(1a') There is a long string.
(1b) Agent(produces, string)
(1b') The string is producing something.
(1c) Product(produces, pitch)
(1c') A pitch is being produced.
(1d) NMod(pitch, low)
(1d') The pitch is low.


Figure 1. Dependency parse tree for example (1)

Breaking the reference answer down into fine-grained facets permits a more focused assessment of the student's response, but a simple yes or no entailment at the facet level still lacks semantic expressiveness with regard to the relation between the student's answer and the reference answer facet in question. For example, did the student contradict the facet or completely fail to address it? Did they express a related concept that indicates a misconception? Did they leave the facet unaddressed? Can you assume that they understand the facet even though they did not express it, (e.g., it was part of the information given in the question)? It is clear that, in addition to breaking the reference answer into fine-grained facets, it is also necessary to break the annotation labels into finer levels in order to specify more clearly the relationship between the student's answer and the reference answer aspect. There are many other representational issues that the system must be able to handle in order to achieve near optimal tutoring, but these two – breaking the reference answer into fine-grained facets and utilizing more expressive annotation labels – are the emphasis of this work.

## 3. Answer Annotation

### 3.1 Corpus

Because most text comprehension problems take root in elementary school during the early years of learning to read and comprehend text, this work focuses on those critical grades. Not yet having interactions with an automated tutoring system, we acquired data gathered from 3rd-6th grade students utilizing the Full Option Science System (FOSS), a proven research-based system that has been in use across the United States for over a decade (Lawrence Hall of Science, 2005). Assessment is a major FOSS research focus, a key component of which is the Assessing Science Knowledge (ASK) project, "designed to define, field test, and validate effective assessment tools and techniques to be used by grade 3–6 classroom teachers to assess, guide, and confirm student learning in science" (Lawrence Hall of Science, 2006).

FOSS includes sixteen diverse science teaching and learning modules (see Table 1) and for each module, the FOSS research team designed a set of summative assessment questions with reference answers. These assessments included multiple-choice questions, fill in the blank questions, and questions requesting drawings, as well as constructed response questions. We reviewed ASK's constructed response questions and selected all of those that were in line with our research goals, which consisted of 287 questions. A representative sample of the questions selected with their reference answers and an example student answer are shown in Table 2.

These questions had expected responses ranging in length from moderately short verb phrases to several sentences. We eliminated fill in the blank questions and questions that we thought were likely to result in short noun phrase answers regardless of the length of the reference answer, assuming these could generally be successfully assessed by most of today's systems and would not benefit from a more fine-grained analysis. Examples of such questions from the Physics of Sound module along with their reference answers and example student responses follow.

Question: *Besides air, what (if anything) can sound travel through?*
Reference Answer: *Sound can also travel through liquids and solids. (Also other gases.)*
Student Answer: *A screen door.*

Question: *Name a property of the sound of a fire engine's siren.*
Reference Answer: *The sound is very loud. OR The sound changes in pitch.*
Student Answer: *Annoying.*

We also eliminated most questions that could not be assessed objectively or that were very open ended. Examples of such constructed response items are:

Question: *Design an investigation to find out a plant's range of tolerance for number of hours of sunlight per day. You can use drawings to help explain your design.*

Question: *Design a way to use carbon printing to find out if two Labrador retrievers have the same paw patterns. Be sure your plan will not be harmful to the dogs.*

Still, there were several moderately open ended questions within the 287 selected. Generally, open ended questions were included if it seemed highly likely that students would address the same points that were included in the reference answer. An example of a question in this category follows.

Question: *What should you do if it appears that an animal is being harmed during an investigation?*
Reference Answer: *Answers will vary. Examples: Be more careful with the animal. Stop the investigation. Change the investigation so it is safer for the animal.*

| Grade | Life Science | Physical Science and Technology | Earth and Space Science | Scientific Reasoning and Technology |
|---|---|---|---|---|
| 3-4 | HB: Human Body<br>ST: Structure of Life | ME: Magnetism & Electricity<br>PS: Physics of Sound | WA: Water<br>EM: Earth Materials | II: Ideas & Inventions<br>MS: Measurement |
| 5-6 | FN: Food & Nutrition<br>EV: Environments | LP: Levers & Pulleys<br>MX: Mixtures & Solutions | SE: Solar Energy<br>LF: Landforms | MD: Models & Designs<br>VB: Variables |

Table 1. FOSS / ASK Learning and Assessment Modules by Area and Grade

| | | |
|---|---|---|
| HB | Q: | Dancers need to be able to point their feet. The tibialis is the major muscle on the front of the leg and the gastrocnemius is the major muscle on the back of the leg. Describe how the muscles in the front and back of the leg work together to make the dancer's foot point. |
| | R: | The muscle in the back of the leg (the gastrocnemius) contracts and the muscle in the front of the leg (the tibialis) relaxes to make the foot point. |
| | A: | The back muscle and the front muscle stretch to help each other pull up the foot. |
| ST | Q: | Why is it important to have more than one shelter in a crayfish habitat with several crayfish? |
| | R: | Crayfish are territorial and will protect their territory. The shelters give them places to hide from other crayfish. [Crayfish prefer the dark and the shelters provide darkness.] |
| | A: | So all the crayfish have room to hide and so they do not fight over them. |
| ME | Q: | Lee has an object he wants to test to see if it is an insulator or a conductor. He is going to use the circuit you see in the picture. Explain how he can use the circuit to test the object. |
| | R: | He should put one of the loose wires on one part of the object and the other loose wire on another part of the object (and see if it completes the circuit). |
| | A: | You can touch one wire on one end and the other on the other side to see if it will run or not. |
| PS | Q: | Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not? |
| | R: | Agree. Vibrations are movements and vibrations produce sound. |
| | A: | I agree with Kate because if you talk in a tube it produce sound in a long tone. And it vibrations and make sound. |
| WA | Q: | Anna spilled half of her cup of water on the kitchen floor. The other half was still in the cup. When she came back hours later, all of the water on the floor had evaporated but most of the water in the cup was still there. (Anna knew that no one had wiped up the water on the floor.) Explain to Anna why the water on the floor had all evaporated but most of the water in the cup had not. |
| | R: | The water on the floor had a much larger surface area than the water in the cup. |
| | A: | Well Anna, in science, I learned that when water is in a more open are, then water evaporates faster. So, since tile and floor don't have any boundaries or wall covering the outside, the water on the floor evaporated faster, but since the water in the cup has boundaries, the water in the cup didn't evaporate as fast. |
| EM | Q: | You can tell if a rock contains calcite by putting it into a cold acid (like vinegar). Describe what you would observe if you did the acid test on a rock that contains this substance. |
| | R: | Many tiny bubbles will rise from the calcite when it comes into contact with cold acid. |
| | A: | You would observe if it was fizzing because calcite has a strong reaction to vinegar. |

Table 2. Sample Qs from FOSS-ASK with their reference answer (R) and an example student answer (A).

We generated a corpus from a random sample of the students' handwritten responses to these questions. ASK was pilot tested in several schools across the United States, with each ASK module typically being tested in two to five schools. Therefore, the students whose answers were transcribed represent a reasonably broad spectrum of the population. The only special transcription instructions were to fix spelling errors (since these would be irrelevant in a spoken dialog environment, the target of this work), but not grammatical errors (which would still be relevant), and to skip blank answers and non-answers similar in nature to *I don't know* (since these are not particularly interesting from the research perspective).

Three test sets were created by 1) withholding all the data from three modules (Environment, Human Body and Water) – resulting in a dataset that can be used to test domain-independent performance, 2) withholding all answers to a subset of questions from each of the other modules (22 questions) – resulting in a dataset that can be used to test question-independent performance, and 3) withholding four answers to each of the remaining questions – resulting in a dataset that can be used to test algorithms intended to handle specific predetermined questions. There are 56 questions, 5,557 student answers, and 47,800 fine-grained facet annotations in the domain-independent test set, comprising approximately 20% of all of the questions utilized and 33% of the total number of facet annotations. There are 22 questions, 997 student answers, and 9,692 facet annotations in the question-independent test set, comprising approximately 8% of all of the questions and 7% of the facet annotations. The third test set spans the remaining 73% of the questions and includes 852 learner responses and 8,700 facet annotations or 6% of all the annotations. This resulted in around 45% of the facet annotations being set aside for testing the learning algorithms, with the remaining 55% (79,719 of 145,911) designated for training and development tuning (7,951 of 15,357 answers).

We selected the three domain-independent test set modules because they appeared to be representative of the entire corpus in terms of difficulty and appropriateness for the types of questions that met our research interests. They were also roughly average sized modules in terms of their number of questions. The items included in the question-independent test set were chosen

randomly, but with two restricting criteria. First, the items were chosen to include at least one question from each module in the training set and to, otherwise, maintain approximately the same question proportions as the training set (the five smallest modules had only one question, the largest had three, and the remaining seven modules had two questions). Second, we did not include questions whose reference answers had significant overlap with questions that would remain in the training data.

In order to maximize the diversity of language and knowledge represented by the training and test datasets, random selection of students was performed at the question level rather than using the same students' answers for all of the questions in a given module. However, in total there were only about 200 children that participated in any individual science module assessment, so there is still moderate overlap in the students from one question to another within a given module. On the other hand, each assessment module was given to a different group of children, so there is no overlap in students between modules.

## 3.2 Annotation

The annotation of student answers consists of two principal steps. First, each reference answer in the corpus, as specified by the ASK research team, was decomposed by hand into its constituent facets. Then each student answer was annotated relative to the facets in the corresponding reference answer to describe whether and how the student addressed those facets. Every student answer was annotated independently by two individuals and a third annotator reviewed the others' labels and made the final decision on each facet's label.

### 3.2.1 Reference Answer Decomposition

The ASK assessments included a reference answer for each of their constructed response questions. These reference answers were broken down into low-level facets, roughly extracted from the relations in a syntactic dependency parse (c.f., Nivre and Scholz, 2004) and a shallow semantic parse (Gildea and Jurafsky, 2002). This decomposition was performed by hand by an undergraduate Linguist and then reviewed for consistency. Since the decomposition is based closely on well established frameworks, dependency parsing and shallow semantic parsing, it was not included in the scope of the experimental research – no formal guidelines were written and the facets were not double annotated to calculate inter-annotator agreement. Generating gold standard referenced answer facets, rather than automatically extracting them, ensured higher quality entailment annotation downstream.

The Physics of Sound reference answers were distilled into their most critical elements. However, minimal changes were made to the remaining answers, since it is desirable for the system to be capable of handling future reference answers written by educators who do not have detailed knowledge of the assessment system, and in the long-term, to handle questions and reference answers generated automatically by the ITS. The most common transformations were to replace nearly all pronouns with their coreferring nouns and to occasionally drop small parts of sentences that were not relevant to the key concepts. The following is a typical example that illustrates each of these modifications in italics.

> Original Reference Answer: *James* should compare the pattern of the pigments on the chromatograms. (If *they* are similar the pens were probably made by the same company.)
>
> Modified Reference Answer: Compare the pattern of the pigments on the chromatograms. If *the chromatograms* are similar the pens were probably made by the same company.

The decomposition of the final reference answers began by determining the dependency parse, following the style of MaltParser (Nivre et al., 2006). This dependency parse was then modified in several ways. Figure 2 shows the standard MaltParser dependency parse and the revised parse for a reference answer fragment that includes several of the issues discussed in the following paragraphs. Example 2 illustrates the decomposition of this same answer fragment into its constituent facets along with their glosses.

(2) The brass ring would not stick to the nail because the ring is not iron.
(2a) NMod(ring, brass)
(2a') The ring is brass.
(2b) Theme_not(stick, ring)
(2b') The ring does not stick.
(2c) Destination_to_not(stick, nail)
(2c') Something does not stick to the nail.
(2d) Be_not(ring, iron)
(2d') The ring is not iron.
(2e) Cause_because(2b-c, 2d)
(2e') 2b and 2c are caused by 2d.

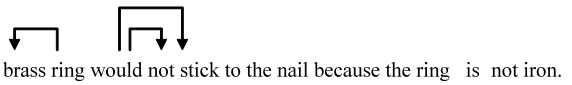


Figure 2. Typical dependency parse revisions to extract reference answer facets

First, wherever a shallow semantic parse would identify a predicate argument structure, we used thematic role labels (c.f., Kipper, Dang and Palmer, 2000) between the predicate and the argument's headword, rather than the MaltParser dependency tags. This also involved, adding new structural dependencies that a typical dependency parser would not generate. For example, in the sentence "As it freezes the water will expand and crack the glass", *water* is not a modifier of *crack* in the typical dependency tree, but it does play the role of Agent in a semantic parse. In a small number of instances, these labels were also attached to noun modifiers, most notably the Location label. For example, given the reference answer fragment *The water on the floor had a much larger surface area*, one of the facets extracted was Location_on(*water, floor*).

The manual parses raised lexical items to governor status when they contextually carried more significant semantics. For example, in the noun phrase *the bunch of leaves*, typically *bunch* is considered the syntactic governor. Whereas, we treat *leaves* as the governor, because it carries more semantics. The parses were also modified to incorporate prepositions, copulas, terms of negation, and similar terms into the dependency type labels (c.f., Lin and Pantel, 2001). This can be seen in the second (revised) reference answer parse in Figure 2, where *to*, *because*, *be*, and *not* were incorporated into the relations of the consolidated dependencies, (e.g., normally *ring is not iron* is parsed as three dependencies, Sub(*is, ring*), VMod(*is, not*), and Prd(*is, iron*), but here they are combined into the single dependency Be_not(*ring, iron*)). When auxiliaries did not contribute much to the semantics of the reference answer, they were not included in the facets extracted.

We refer to facets that express relations between higher-level propositions as inter-propositional facets. An example of such a facet is (2e) above, connecting the proposition *the brass ring did not stick to the nail* to the proposition *the ring is not iron*. In addition to specifying the headwords of inter-propositional facets (*stick* and *is*, in 2e), we also indicate up to two key facets from each of the propositions that the relation is connecting (b, c, and d in example 2). Reference answer facets that are assumed to be understood by the learner a priori, (e.g., because they are part of the question), are also marked to indicate this.

There were a total of 2877 reference answer facets, resulting in a mean of 10 facets per question (median of 8 facets). Table 3 shows a high-level break down of the reference answer facets. Facets that were assumed to be understood a priori by students accounted for 33% of all facets and inter-propositional facets accounted for 11%. The experiments in automated annotation of student answers (section 4) focus on the facets that are not assumed to be understood a priori (67% of all facets); of these, 12% are inter-propositional.

| Category | Freq. | Freq./Q | % of Total | % (not) assumed |
|---|---|---|---|---|
| All facets | 2877 | 10.0 | 100 | |
| Assumed | 949 | 3.3 | 33 | |
| Not assumed | 1928 | 6.7 | 67 | |
| Inter-propositional | 326 | 1.1 | 11 | |
| Simple | 2551 | 8.9 | 89 | |
| Inter-propositional assumed | 100 | 0.3 | 3 | 11 |
| Simple assumed | 849 | 3.0 | 30 | 89 |
| Inter-propositional not assumed | 226 | 0.8 | 8 | 12 |
| Simple not assumed | 1702 | 5.9 | 59 | 88 |

Table 3. Frequency of reference facets by category

A total of 35 different facet relation types were utilized (see Table 4). The majority, 21, are VerbNet thematic roles. Direction, Manner, and Purpose are PropBank adjunctive argument labels (Palmer et al., 2005). Quantifier, Means, Cause-to-Know, copulas and similar verbs (e.g., be, become, do, and have) were also used as facet relation types. Finally, anything that did not fit into the above categories retained its dependency parse type: VMod (Verb Modifier), NMod (Noun Modifier), AMod (Adjective or Adverb Modifier), and Root (Root was used when a single word in the answer, typically yes, no, agree, disagree, A-D, or a number, stood alone without a significant relation to the remainder of the reference answer; this occurred only 23 times, accounting for fewer than 1% of the reference answer facets). The seven highest frequency relations are NMod, Theme, Cause, Be, Patient, AMod, and Location, which together account for 70% of the reference answer facet relations.

| VerbNet Role | Not Asmd | Asmd | Ttl | Other Roles | Not Asmd | Asmd | Ttl |
|---|---|---|---|---|---|---|---|
| Actor | 0 | 1 | 1 | *PropBk Adjs* | | | |
| Agent | 22 | 35 | 57 | Direction | 15 | 4 | 19 |
| Attribute | 12 | 0 | 12 | Manner | 37 | 6 | 43 |
| Beneficiary | 3 | 0 | 3 | Purpose | 2 | 2 | 4 |
| Cause | 164 | 80 | 244 | | | | |
| Destination | 57 | 17 | 74 | *Misc. Types* | | | |
| Experiencer | 8 | 7 | 15 | Cause-know | 15 | 12 | 27 |
| Extent | 23 | 10 | 33 | Means | 51 | 27 | 78 |
| Instrument | 13 | 7 | 20 | Quantifier | 61 | 26 | 87 |
| Location | 84 | 43 | 127 | | | | |
| Material | 26 | 9 | 35 | *Special Vbs* | | | |
| Patient | 93 | 44 | 137 | Be | 140 | 50 | 190 |
| Predicate | 31 | 5 | 36 | Become | 7 | 2 | 9 |
| Product | 47 | 15 | 62 | Do | 1 | 0 | 1 |
| Recipient | 14 | 6 | 20 | Have | 23 | 25 | 48 |
| Source | 11 | 5 | 16 | | | | |
| Stimulus | 9 | 2 | 11 | *Dependency* | | | |
| Theme | 279 | 127 | 406 | AMod | 112 | 24 | 136 |
| Time | 65 | 20 | 85 | NMod | 447 | 322 | 769 |
| Topic | 11 | 6 | 17 | Root | 23 | 0 | 23 |
| Value | 3 | 1 | 4 | VMod | 19 | 9 | 28 |

Table 4. Reference answer facet types and frequencies

### 3.2.2 Annotation Guidelines

The answer assessment annotation described in this chapter is intended to be a step toward specifying the detailed semantic understanding of a student's answer that is required for an ITS to interact as effectively as possible with a learner. With that goal in mind, annotators were asked to consider and annotate according to what they would want to know about the student's answer if they were the tutor. However, we only annotate a student's answer relative to the constituent facets of the reference answer. If the student also discusses concepts not addressed in the reference answer, those points are not annotated regardless of their quality or accuracy.

After analyzing much of the Physics of Sound data, we settled on the eight annotation labels noted in Table 5 (Nielsen and Ward, 2007). Descriptions of where each annotation label applies and some of the most common annotation issues were detailed with several examples in the guidelines and are summarized below.

Example 3 shows a question and a fragment of its reference answer broken down into its constituent facets with an indication of whether the facet is assumed to be understood a priori. A corresponding student answer is shown in (4) along with its final annotation in 3a'-c'. It is assumed that the student understands that the pitch is higher a priori (reference answer facet 3b), since this is given in the question (… *Write a note to David to tell him*

*why the pitch gets higher rather than lower*) and similarly it is assumed that the student will be explaining what has the causal effect of producing this higher pitch (facet 3c). Therefore, unless the student explicitly addresses these facets they are labeled *Assumed*.

| |
|---|
| **Assumed**: Reference answer facets that are assumed to be understood a priori, most often based on the question |
| **Expressed**: Any reference answer facet directly expressed or inferred by simple reasoning |
| **Inferred**: Reference answer facets inferred by pragmatics or nontrivial logical reasoning |
| **Contra-Expr**: Reference answer facets directly contradicted by negation, antonymous expressions and their paraphrases |
| **Contra-Infr**: Reference answer facets contradicted by pragmatics or complex reasoning |
| **Self-Contra**: Reference answer facets that are both contradicted and implied (self contradictions) |
| **Diff-Arg**: The core relation is expressed, but it has a different modifier or argument |
| **Unaddressed**: Reference answer facets that are not addressed at all by the student's answer |

Table 5. Facet Annotation Labels

(3) Question: After playing the FOSS-ulele, David wrote his results in his lab notebook:
*I'm confused. When I pull down and tighten the string on the FOSS-ulele, then pluck the string, the pitch sounds HIGHER than it did before. But aren't I making the string longer when I pull the string? I thought a longer length produced a LOWER pitch. What's going on here?*
What is causing the pitch to be higher? Write a note to David to tell him why the pitch gets higher rather than lower.
Reference Answer: The string is tighter, so the pitch is higher.
(3a) Be(string, tighter), ---
(3b) Be(pitch, higher), Assumed
(3c) Cause(3b, 3a), Assumed

(4) David this is why because you don't listen to your teacher. If the string is long, the pitch will be high.
(3a') Be(string, tighter), Diff-Arg
(3b') Be(pitch, higher), Expressed
(3c') Cause(3b', 3a'), Expressed

Since the student does not contradict the fact that the string is tighter (the string can be both longer and tighter), we do not label this facet as *Contradicted*. If the student's response did not mention anything about either the *string* or *tightness*, we would annotate reference answer facet 3a' as *Unaddressed*. However, the student did discuss a property of the string, *the string is long*. This parallels the reference answer facet Be(*string, tighter*) with the exception of a different argument to the *Be* relation, resulting in the annotation *Diff-Arg*. This indicates to the tutor that the student expressed a related concept, but one which neither implies that they understand the facet nor that they explicitly hold a contradictory belief. Often, this indicates the student has a misconception. For example, when asked about an effect on pitch, many students say things like the *pitch gets louder*, rather than higher or lower, which implies a misconception involving their understanding of pitch and volume. In this case, the Diff-Arg label can help focus the tutor on correcting this misconception. Facet 3c', expressing the causal relation between 3a' and 3b', is labeled *Expressed*, since the student did express a causal relation between the concepts aligned with 3a' and 3b'. The tutor then knows that the student was on track in regard to attempting to express the desired causal relation and the tutor need only deal with the fact that the cause given was incorrect.

The *Self-Contra* annotation is used in cases like the response in example 5, where the student simultaneously expresses the contradictory notions that the string is tighter and that there is less tension.

(5) The string is tighter, so there is less tension so the pitch gets higher.
(3a") Be(string, tighter), *Self-Contra*
(3b") Be(pitch, higher), Expressed
(3c") Cause(3b", 3a"), Expressed

Example 6 illustrates a case where a student's answer is labeled *Inferred*. In this case, the decision requires pragmatic inferences, applying the Gricean maxims of Relation, be relevant – why would the student mention vibrations if they did not know they were a form of movement – and Quantity, do not make your contribution more informative than is required (Grice, 1975).

(6) Question: Kate said: "An object has to move to produce sound." Do you agree with her? Why or why not?
Reference Answer: "Agree. Vibrations are movements and vibrations produce sound."
Student Answer: Yes because it has to vibrate to make sounds.
(6a) Root(root, agree), Expressed
(6b) Be(vibration, movement), *Inferred*
(6c) Agent(produce, vibrations), Expressed
(6d) Product(produce, sound), Expressed

There is no compelling reason from the perspective of the automated tutoring system to differentiate between Expressed, Inferred and Assumed facets, since in each case the tutor can assume that the student understands the concepts involved. However, from the systems development perspective there are three primary reasons for differentiating between these facets and similarly between facets that are contradicted by inference versus by more explicit expressions. The first reason is that most systems today cannot hope to detect very many pragmatic inferences, which are the main source of the Inferred and Assumed labels, and including these in the training data can sometimes confuse learning algorithms resulting in worse performance. Having separate labels allows one to remove the more difficult inferences from the training data, thus eliminating this issue. The second rationale is that systems hoping to handle both types of inference might more easily learn to discriminate between these opposing classifications if the classes are distinguished (for algorithms where this is not the case, the classes can easily be combined automatically). Similarly, this allows the possibility of training separate classifiers to handle the different forms of inference. The third reason for separate labels is that it can facilitate system evaluation, including the comparison of various

techniques and the effect of individual features – one can assess separately whether a technique or feature had a positive or negative impact on the Inferred facets or on the Expressed facets.

Annotators were all college students, ranging from first year undergraduates to graduate students and came from a variety of departments including Education, Linguistics, Computer Science, and Cognitive Science. In total, seven annotators were involved over the course of the project. Generally, the same annotator performed the entire first, second, or adjudication tagging for all of the questions in a given science module to reduce the learning curve.

### 3.3 Inter-Annotator Agreement Results

We report results under three groupings: (1) *All-Labels*, where all labels are left separate, (2) *Tutor-Labels*, consisting of the five labels that will be used by the automated tutor, where Expressed, Inferred and Assumed are combined into a single *Understood* class (i.e., the annotator believes the student understands the facet) and Contra-Expr and Contra-Infr are replaced with *Contradicted* (i.e., the annotator believes the student holds a view contradictory to the reference answer facet), and (3) *Yes-No*, which is a binary decision, Understood versus all other labels. The Tutor-Labels grouping will likely be used by the ITS, since it is relatively unimportant to differentiate between the types of inference required in determining that the student understands a reference answer facet.

We calculate inter-annotator agreement and Cohen's Kappa statistic (Cohen, 1960) based on all 16 of the science modules, totaling 142,451 total facet annotations[2]. Agreement on the Tutor-Labels is 86.2%, with a Kappa statistic of 0.728, corresponding with substantial agreement. Agreement is 78.4% on All-Labels and 88.0% on the binary Yes-No decision. These too have Kappa statistics in the range of substantial agreement (see Table 7 for details).

| Label Grouping | ITA % | Kappa |
|---|---|---|
| All-Labels | 78.4% | 0.704 |
| Tutor-Labels | 86.2% | 0.728 |
| Yes-No | 88.0% | 0.752 |

Table 6. Inter-annotator agreement by label groupings

The distribution of facet annotations is shown in Table 7. The most frequent fine-grained label is Unaddressed, at 36.0%, and the majority, 61.1%, of the Tutor-Labels indicate the student understood the reference answer facet. An analysis of the inter-annotator confusion matrix indicates that the most probable disagreement is between Inferred and Unaddressed, representing 39% of all the disagreements. The next most likely disagreements are between Expressed and the other Understood labels (Inferred and Assumed), comprising 35% of the disagreements. Confusion between Expressed and Unaddressed is also considerable, representing 10% of the annotator disagreements.

[2] Part of the data used during annotator training was not double annotated and thus is not included in Table 6.

| Label | Count | % | Count | % |
|---|---|---|---|---|
| Expressed | 31,555 | 21.6 | 89,105 | 61.1 |
| Inferred | 20,474 | 14.0 | | |
| Assumed | 37,076 | 25.4 | | |
| Contra-Expr | 1,426 | 1.0 | 2,482 | 1.7 |
| Contra-Infr | 1056 | 0.7 | | |
| Self-Contra | 86 | 0.1 | 86 | 0.1 |
| Diff-Arg | 1,780 | 1.2 | 1,780 | 1.2 |
| Unaddressed | 52,458 | 36.0 | 52,458 | 36.0 |

Table 7. Distribution of classifications (145,911 facets)

## 4. Results of Automated Classification

We implemented a machine learning based classifier following Dagan et al. (2005) and Nielsen et al. (2006). (See (Nielsen et al., 2008) for details regarding the system.) A high level description of the system classification procedure follows. We start with the hand generated reference answer facets. We generate automatic parses for the reference answers and the student answers and automatically modify these parses per our desired representation. Then for each reference answer facet, we extract features indicative of the student's understanding of that facet. Finally, we train a machine learning classifier on the training data and use it to classify unseen test examples, assigning a separate Tutor-Label for each reference answer facet to indicate the student's understanding of that reference answer facet.

Training and testing excluded facets that were Assumed to be understood a priori. At the time of training and evaluation, we had 54,967 facet annotations in the training set, 30,514 examples in the Unseen Modules test set, 6,699 examples in the Unseen Questions test set and 3,159 examples in the Unseen Answers test set. Table 6 shows the classifier's accuracy (percent correctly tagged) in 10-fold cross validation on the training set as well as on each of our test sets. The row labeled Unseen Answers presents the accuracy when classifying different answers to the same questions that generated the training set answers. Unseen Questions provides the accuracy of classifying answers to questions not used in the training set and Unseen Modules shows the accuracy on the domain-independent test data collected from very different science modules than were used for training. The first two columns show the simple baseline accuracies of 1) a classifier that always outputs the most frequent class in the training set – Unaddressed, and 2) a lexical decision that outputs Understood if both the governing term and the modifier are present in the learner's answer and outputs Unaddressed otherwise. The *ML System* column represents the accuracy of the full classifier in predicting the five labels that will drive the system dialogue: Understood (Expressed and Inferred), Contradicted (Contra-Expr and Contra-Infr), Self-Contra, Diff-Arg, and Unaddressed.

| | Majority Class | Lexical Baseline | ML System |
|---|---|---|---|
| Training 10x-CV | 54.6 | 59.7 | 77.1 |
| Unseen Answers | 51.1 | 56.1 | 75.5 |
| Unseen Questions | 58.4 | 63.4 | 66.5 |
| Unseen Modules | 53.4 | 62.9 | 68.8 |

Table 8. Classifier Accuracy on Tutor-Labels

## 5. Summary

The goal of our fine-grained classification is to enable more effective tutoring dialog management. The additional labels facilitate understanding the type of mismatch between the reference answer and the student's answer. Breaking the reference answer down into low-level facets enables the tutor to provide feedback relevant specifically to the appropriate facet of the reference answer. In the question answering domain, this facet-based classification would allow systems to accumulate entailing evidence from a variety of corroborating sources and incorporate answer details that might not be found in any single sentence. Similarly, in multi-document summarization, entailment at the facet level could help systems recognize or verify important aspects of a topic. This fine-grained classification can also facilitate more directed user feedback outside of the tutoring domain. For example, both the additional classifications and the break down of facets can be used to justify system decisions.

The corpus described in this paper, which is publicly available for research purposes[3], was annotated with substantial inter-annotator agreement at 86.2%, (Kappa=0.728) and represents a substantial contribution to the entailment community, including 145,911 facet entailment annotations. By contrast, three years of RTE challenge data comprise fewer than 4,600 entailment annotations. More importantly, this is the only corpus that provides entailment information at the fine-grained level described in this paper.

This is a new task and new dataset and the results are very promising. Classification according to the Tutor-Labels is 24.4%, 8.1%, and 15.4% over the most frequent class baseline for Unseen Answers, Questions, and Modules respectively. These results demonstrate that the task is feasible and we believe will become an effective component in entailment applications. We are currently working toward integrating the classifier into an intelligent tutoring system.

## 6. Acknowledgements

This work was partially funded by Award Number 0551723 from the National Science Foundation.

## 7. References

Callear, D., Jerrams-Smith, J. and Soh, V. (2001). CAA of short non-MCQ answers. In *5th Intl CAA*.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational & Psych Measurement*. 20:37-46.

Dagan, I., Glickman, O. and Magnini, B.. (2005). The PASCAL Recognizing Textual Entailment Challenge. In *1st RTE Challenge Workshop*.

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*:3, 245–288.

Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M. and Olde, B. (2001). AutoTutor: An Intelligent Tutor and Conversational Tutoring Scaffold. In *10th ICAI in Education*, 47-49.

Grice, H. Paul. (1975). Logic and conversation. In P Cole and J Morgan (Eds), *Syntax and Semantics, Vol 3, Speech Acts*, 43–58. Academic Press.

Jordan, P.W., Makatchev, M. and VanLehn, K. (2004). Combining competing language understanding approaches in an intelligent tutoring system. In *7th ITS*.

Kipper, K., Dang, H. and Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI17th NCAI*

Lawrence Hall of Science (2005) Full Option Science System (FOSS), University of California at Berkeley, Delta Education, Nashua, NH.

Lawrence Hall of Science (2006) Assessing Science Knowledge (ASK), University of California at Berkeley, NSF-0242510

Leacock, C. (2004). Scoring free-response automatically: A case study of a large-scale Assessment. *Examens, 1*(3).

Lin, D. and Pantel, P. (2001). Discovery of inference rules for Question Answering. In *Natural Language Engineering, 7*(4):343-360.

Mitchell, T. Aldridge, N. and Broomhead, P. (2003). Computerized marking of short-answer free-text responses. In *29th IAEA*.

Nielsen, R.D., Ward, W. and Martin, J.H. (2006). Toward dependency path based entailment. In *Proc. of the second PASCAL RTE challenge workshop*

Nielsen, R.D., Ward, W. (2007). A corpus of fine-grained entailment relations. In *Proc. of the ACL workshop on Textual Entailment and Paraphrasing*.

Nielsen, R.D., Ward, W., and Martin, J.H. (2008). Learning to Assess Low-level Conceptual Understanding. In Proc. FLAIRS

Nivre, J., Hall, J., Nilsson, J., Eryigit, G. and Marinov, S. (2006). Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL)*.

Nivre, J. and Scholz, M. (2004). Deterministic Dependency Parsing of English Text. In *Proc COLING*.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. In *Computational Linguistics*.

Peters, S., Bratt, E.O., Clark, B., Pon-Barry, H. and Schultz, K. (2004). Intelligent Systems for Training Damage Control Assistants. In *Proc. of ITSE*.

Pulman S.G. and Sukkarieh J.Z. (2005). Automatic Short Answer Marking. *ACL WS Bldg Ed Apps using NLP*.

Roll, I, Baker, R., Aleven, V., McLaren, B. and Koedinger, K. (2005). Modeling Students' Metacognitive Errors in Two Intelligent Tutoring Systems. In *UM* 379–388

Rosé, P. Roque, A., Bhembe, D. & VanLehn, K. (2003). A hybrid text classification approach for analysis of student essays. In *Bldg Ed Apps using NLP*

VanLehn, K., Lynch, C., Schulze, K. Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A. and Wintersgill, M. (2005). The Andes physics tutoring system: Five years of evaluations. In *12th ICAI in Ed*

[3] Contact the first author to acquire the corpus.