Automatic Evaluation of Learner Self-Explanations and Erroneous Responses for Dialogue-Based ITSs

Blair Lehman[1], Caitlin Mills[2], Sidney D'Mello[2], and Art Graesser[1]

[1]Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152
{balehman|a-graesser}@memphis.edu

[2]Department of Psychology, University of Notre Dame, South Bend, IN 46556
{cmills4|sdmello}@nd.edu

Abstract. Self-explanations (SE) are an effective method to promote learning because they can help students identify gaps and inconsistencies in their knowledge and revise their faulty mental models. Given this potential, it is beneficial for intelligent tutoring systems (ITS) to promote SEs and adaptively respond based on SE quality. We developed and evaluated classification models using combinations of SE content (e.g., inverse weighted word-overlap) and contextual cues (e.g., SE response time, topic being discussed). SEs were coded based on correctness and presence of different types of errors. We achieved some success at classifying SE quality using SE content and context. For correct vs. incorrect discrimination, context-based features were more effective, whereas content-based features were more effective when classifying different types of errors. Implications for automatic assessment of learner SEs by ITSs are discussed.

Keywords: self-explanations, automatic scoring, adaptive responses, ITSs, natural language understanding

1 Introduction

Learning is a complex process that involves both the acquisition of new knowledge and integration of new content with existing knowledge. This task can be especially difficult when learners' mental models are rife with gaps, inconsistencies, and misconceptions. One method to facilitate the learning process is to have instructors provide explanations and guidance. Another method is to allow learners to construct and refine their own mental models. The latter method represents a more active form of knowledge construction. This type of active knowledge construction, in which learners are encouraged to engage in a form of self-instruction [1], can be contrasted with more shallow learning that involves the mere accumulation of facts [2-3].

Self-instruction can be completed through a number of learning activities; one such activity is self-explanation. Self-explanations (SE) are a representation of the learner's current knowledge about a concept and involve making inferences as well as integrating new information into existing knowledge structures [4]. SEs can also

facilitate learning by causing learners to realize where gaps or inconsistencies exist in their knowledge [5-6]. The impact of SEs on learning can be especially strong when learners are required to apply skills to new situations [5, 7].

The value of SEs as a means to diagnose learner knowledge and facilitate learning has been acknowledged for some time. Many studies have taken advantage of the SE effect (e.g., [5, 8, 9]). For example, Chi et al. [5] had learners study example problems on Newtonian physics and engage in a talk-aloud while studying. They found that higher achieving learners generated SEs at each step of the example problem while working to create a more refined understanding of the concept. Less successful learners, on the other hand, did not generate their own SEs while learning.

The benefits of SEs have also been studied in the context of intelligent tutoring systems (ITS). Many ITSs incorporate SEs as part of the learning process and some even train learners to become more adept self-explainers [6-7, 10-12]. iSTART, for example, is an ITS that provides learners with SE and reading strategy training [12]. By providing learners with examples of high quality SEs, practice generating SEs, and additional reading strategies, iSTART is able to increase learners' reading comprehension skills [13].

In addition to promoting SE use and training learners to generate higher quality SEs, ITSs must also be capable of evaluating the quality of learner-generated SEs. If an ITS can provide learners with opportunities to self-explain and automatically assesses the quality of their SEs, the ITS can adaptively respond to any gaps in the learner's knowledge and begin to correct problematic misconceptions.

The process of understanding natural language contributions from learners, however, is not a trivial task because the responses are often short, conversational, fragmented, and syntactically incorrect. In one study, Williams and D'Mello [14] used linguistic properties to assess the quality of learner responses during expert human tutoring sessions. The Linguistic Inquiry Word Count [15] was used to classify answers as correct, partially-correct, vague, or error-ridden. Although this approach did not use any content-dependent words, they were able to correctly classify 45.2% of learner responses.

Other studies have used a more content-dependent approach for assessing learner contributions. Litman, Moore, Dzikovska, and Farrow [16] used content word matching to analyze corpora from tutoring sessions with an ITS and human tutors. Use of a domain-specific glossary yielded some success; however, approximately half of the content words in learner responses were misclassified. In a series of studies, Graesser, Penumatsa, Ventura, Cai, and Hu [17] made use of Latent Semantic Analysis (LSA) [18] to model learner knowledge during interactions with an ITS. LSA is a method to semantically compare two texts using a bag of words approach and dimensionality reduction techniques. By comparing learner responses to expectations (ideal responses) and common misconceptions, they were able to model learner knowledge at a level that was comparable to unskilled human tutors.

Research on natural language understanding (NLU) techniques to assess learner responses has also revealed that a combination of algorithms may be an effective method for diagnosing learner knowledge. Aleven, Popescu, and Koedinger [19] used the combination of a geometry knowledge base (e.g., keywords, ideal responses) and a statistical text classifier (NaïveBayes). The knowledge base incorporated hierarchical ordering for comparisons of learner responses to correct or partially-

correct example responses. When only the knowledge base was used to discriminate between correct and incorrect learner responses, 59.5% of responses were correctly classified [20]. However, when the classification model included both the knowledge base and statistical classifier, classification improved to 61% [21]. The negligible increase, when the statistical classifier was included (59.5% vs. 61%), was attributed to the large number of potential classifications for each SE (167 labels). When semantic similarity between labels, or types of error-ridden answers, was taken into account and reduced the number of potential labels, accuracy greatly increased to 81%.

Rus, McCarthy, Lintean, Graesser, and McNamara [22] examined seven algorithms to assess the quality of learner SEs from iSTART interactions. iSTART presents learners with a text and then asks them to explain the text in their own words. The algorithms were either word-based, syntactic, or a combination of word and syntactic information. Word-based algorithms assessed word-overlap between learner SEs and the original text. Seventy-four percent of paraphrase SEs were correctly classified via a combination of the entailment index [23], synonymy index, word-overlap, and LSA (see [22] for details).

Past research on automatic classification of learner contributions has focused on the response content (i.e., the words in the response), while context from the learning session has largely been ignored. In the present paper we attempt to expand upon these past results by augmenting a semantic analysis of the response content with information about the context surrounding the response. Similar to past research, we test a model that uses a weighted word-overlap algorithm as the predictive feature (SE *Content* model). We build on past research by testing a *Context* model that incorporates features of the response characteristics (e.g., SE response time) and larger learning context (e.g., order of topic presentation, prior performance in the learning session). We compare the individual models to a *Combined* model (Content + Context). Finally, taking a somewhat different approach, we tested a *Word-Based* model that exclusively relies on the words in SEs as predictive features without a knowledge-based model of correct and incorrect answers. The models were tested on a corpus of learner SEs collected from a previous study involving tutorial sessions on scientific reasoning topics.

2 Method

2.1 Participants

Participants were 76 undergraduate students from a mid-south university in the US who participated for course credit. Participants completed four learning sessions, one on each of the scientific reasoning topics (experimenter bias, control group, random assignment, replication). Order of topics was counterbalanced across participants.

2.2 Learning Sessions

The learning sessions consisted of the human learners engaging in a trialogue with two animated pedagogical agents (tutor agent and peer student agent) while learning about scientific reasoning topics [24]. The trialogues consisted of learners discussing the scientific merits of sample research studies with the tutor and peer agents. All studies had one subtle flaw and the learner had to identify this flaw by reflecting and deliberating on the quality of the methodologies of the studies.

The excerpt in Table 1 is an example trialogue between the two agents and the human learner. In this excerpt, the tutor agent (Dr. Williams) and the peer agent (Chris) are discussing a flawed study with the human learner (Bob). The study is flawed because the replication study used the same population as the original study.

Table 1. Excerpt of trialogue

Turn	Speaker	Dialogue Move
A study showed that people who drank non-alcoholic beer can feel drunk. A professor ran this study with his class. He randomly assigned 50 students to drink the real beer and 50 the non-alcoholic beer. **The professor was skeptical, so he ran a replication study two weeks later with the same method and same class**. He found the same results again. So he concluded that people can feel just as drunk from non-alcoholic beer. *<Research Study>*		
1	Dr. Williams	I think we all need to go over the experiment some more. *<Advance dialogue>*
2	Dr. Williams	Alright, so I think that the real problem here is with the second experiment. *<Assert information>*
3	Chris	I disagree with you, Dr. Williams. I think the replication was done right. *<Assert information>*
4	Dr. Williams	Bob, was this a good or bad replication? Please type good or bad. *<Forced-choice question>*
5	Bob	bad *<Response>*
6	Dr. Williams	**Tell me more about your thoughts on this Bob. *<Pump for SE>***
7	Bob	**How can feel drunk drinking a non-alcoholic drink unless you had a alcoholic drink before taking the experiment. Its faulty and doesn't make sense. *<Student SE>***

Each learning session began with learners reading the research study and then beginning a trialogue with the agents. The discussion involved five trials. All trials consisted of the tutor (turn 2) and peer (turn 3) agents asserting their opinions, prompting the learner to intervene (turn 4), and obtaining the learner's response (turn 5). Learner responses were either correct or incorrect. Trials 3 and 4 also included pumps (turn 6) that required learner SEs (turn 7). Learners did not receive feedback on SE quality; the agents simply acknowledged learners' contributions (e.g., "Alright," "Okay"). This cycle was repeated in each trial, with each trial becoming more specific about the scientific merits of the study. The present paper will focus on Trials 3 and 4 because learners were asked to self-explain during these trials.

2.3 Procedure

Learners were tested individually over a two-hour session. First, learners signed an informed consent and completed the pretest. Next, learners read a short introduction on research methods. Learners then completed four learning sessions, one on each scientific reasoning topic. Finally, learners completed the posttest and were fully debriefed. Pretest and posttest data is not relevant to the present analyses and will not be discussed any further.

2.4 Self-Explanation Coding

A total of 608 learner SEs were obtained from the learning sessions. Two human-raters coded the SEs as correct, partially-correct, or incorrect. A subset of the corpus was first coded to compute reliability (*kappa* = .842). The corpus was then divided evenly between the raters for coding. For the current analyses, partially-correct and incorrect SEs were collapsed into one category (incorrect) because there were very few instances of partially-correct SEs (8.72%). This yielded 36% correct responses and 64% incorrect responses.

Incorrect SEs were further coded for types of error-ridden reasoning. Learner SEs could be rated as Correct, Error Type 1, Error Type 2, Error Type 3, Unclassified, or Frozen Expression. Incorrect learner SEs that did not fit into one of the error type categories were grouped as *Unclassified*. Frozen expressions, SEs unrelated to the topic, were not included in the current analyses because a speech act classifier that can accurately identify these utterances has already been developed [25].

Table 2 shows an example of a correct response, different error types, and a frozen expression. Error types were unique to each scientific reasoning topic and trial. Errors could vary from focusing on superficial features of the study rather than methodological issues (see Error Type 2) to complete misunderstandings of the concept being discussed (see Error Type 1).

Table 2. Examples of SE response types for Trial 3 of the replication topic

Response Type	Example
Correct Answer	It was bad since the study used the same people to replicate the study. Different people should have been used so the accuracy of the data could have been confirmed more firmly.
Error Type 1	I think that it was a good replication of the first study; however, I do not think that the first study was executed properly.
Error Type 2	How can feel drunk drinking a non-alcoholic drink unless you had a alcoholic drink before. It doesn't make sense.
Error Type 3	The professor was careful to conduct random assignment. That helps to make it a good replication. And he used the same people.
Unclassified	It was conducted well but the longevity of the study could not make it very accurate.
Frozen Expression	I don’t know.

2.5 Semantic Matching

In order to evaluate the semantic quality of learner SEs, we first needed to create expected responses and expected errors. Prototypical correct responses and prototypical erroneous responses (for each error type) were created by a content expert (see Table 2 for an example). Prototypical correct and erroneous responses were unique to each of the eight individual questions (4 topics x 2 trials).

Learner SEs were compared to prototypical correct and erroneous responses using an inverse word frequency weighted overlap (IWFWO) algorithm. The IWFWO algorithm is a word-matching algorithm in which each overlapped word is weighted on a scale from 0 to 1, relative to its inverse frequency in the English language using the CELEX corpus [26]. The inverse frequency allows for higher weighting of lower frequency, more contextually relevant words (e.g., replication, bias), while higher frequency words (e.g., and, but) are given a lower weighting. Comparisons resulted in a match score between 0 and 1 (1 = perfect similarity).

3 Results and Discussion

3.1 Content, Context, and Combined Models

We tested three models to determine which SE features were most diagnostic of SE quality. The *Content Model* included the IWFWO match score (either to the prototypical correct or error type SE based on the classification task) and the number of words in the SE. The *Context Model* included SE response time, performance (correct or incorrect) and response time on the forced-choice question prior to the SE (see turn 4 in Table 1), and the order of topic presentation (e.g., first, second). These contextual features were selected because they are already logged by the learning environment and would not require additional processing for future SE classification. Finally, there was also a *Combined* model, which combined features from the two individual models.

Four classification algorithms from WEKA [27] were used to build and evaluate the models: NaïveBayes, IBk (nearest neighbor with $k = 10$), j48, and LogitBoost. The majority class algorithm (ZeroR) that classifies all SEs to the most prevalent group was used as the baseline comparison. Each algorithm was evaluated using 10-fold cross-validation. Two separate classification tasks were performed. The first task consisted of making a simple correct vs. incorrect discrimination, while the second task performed a fine-grained discrimination in terms of specific error types.

SEs were separated into eight groups based on scientific reasoning topic and trial. After removing frozen expression responses, there was an average of 71.9 responses per group (*SD* = 2.42; *Range* 69 to 75). The algorithms were evaluated on each SE group for both classification tasks. For each SE group the best algorithm (i.e., one out of the four algorithms that yielded the best performance) was selected. The best classification results were averaged across SE groups and constituted the Content, Context, and Combined models. Table 3 shows the results obtained for each classification task averaged across the eight groups.

Table 3. Mean (SD) classification performance across groups

	Correct-Incorrect		Error Type	
Model	**Accuracy (%)**	**Kappa**	**Accuracy (%)**	**Kappa**
Baseline	64.6 (9.45)	.000 (.000)	43.3 (6.93)	.000 (.000)
Content	69.5 (6.74)	.248 (.080)	**67.6 (4.44)**	**.501 (.108)**
Context	**74.0 (4.08)**	**.335 (.160)**	50.3 (8.44)	.231 (.095)
Combined	74.3 (3.92)	.347 (.160)	67.4 (6.54)	.510 (.103)

We note that the Context model (74.0%) was the most successful for segregating correct from incorrect responses. Both the Content, $t(7) = 2.40$, $p < .05$, and Context models, $t(7) = 4.29$, $p < .01$, performed significantly better than the Baseline model. The Context model also significantly outperformed the Content model for correct-incorrect discriminations, $t(7) = 2.39$, $p < .05$. Both individual models outperformed the Baseline model for error type discriminations (Content: $t(7) = 8.02$, $p < .01$; Context: $t(7) = 2.69$, $p < .05$). However, it was the Content model that performed best for error discrimination (67.6%). Interestingly, the Content model was twice as more effective for error type classifications than the Context model, $t(7) = 4.70$, $p < .01$. Indeed, these models were differentially effective for different classification tasks.

When comparing correct and incorrect SEs, we found that learners with correct SEs took longer to self-explain, $t(14) = 3.14$, $p = .01$, and responded more accurately to the forced-choice question prior to self-explaining, $t(14) = 2.30$, $p < .05$. This suggests that learners who responded correctly took more time to thoughtfully construct a response. For erroneous SEs, error types only differed on match to the prototypical erroneous responses, $F(3) = 20.2$, $p < .01$, which is what could be expected. Furthermore, SEs that were grouped as *unclassified* had lower match scores to the prototypical erroneous responses.

Comparisons of the Combined model to the individual models were also quite informative. Combined models for both discrimination tasks outperformed the Baseline models (correct-incorrect: $t(7) = 2.86$, $p < .05$; error type: $t(7) = 8.26$, $p < .01$). However, the Combined model did not yield any noticeable improvements over the best performing individual model for either the correct vs. incorrect or error discrimination task ($p's > .05$). The negligible improvement by the Combined models suggests that it may be beneficial for systems to not conduct a full classification model initially, but rather allot these resources only when needed. For example, if an SE is classified as correct, it is not necessary to conduct a full classification model and analyze the actual content of the SE.

3.2 Word-Based Models

We also attempted to classify SEs with only the words in responses as features. This was accomplished using the StringToWordVector package in WEKA to transform text strings (words) into numerical input using *tf-idf* (term frequency-inverse

document frequency) weighting. The tf-idf weighting allows less frequent, more content-rich words to have higher weightings.

The same four classifiers were used to train the models and they were tested with ten-fold cross-validation. As in the previous analyses, SEs were separated by scientific reasoning topic and trial for classification. The best classifier for each individual SE group was then selected. The average classification accuracy (across the eight groups) for the correct vs. incorrect was 71.1% (*SD* = 8.45) with a kappa of .282 (*SD* = .178). For error discrimination, the average accuracy was 58.1% (*SD* = 9.30) with a kappa of .352 (*SD* = .119). The word-based models performed significantly better than the Baseline model for both discrimination tasks (correct vs. incorrect: $t(7) = 2.10, p < .1$; error type: $t(7) = 5.43, p < .01$).

These results suggest that while it is possible to classify SEs on the basis of words alone, the resultant models were less effective than the Content model (67.6% accuracy) for error classification. However, the word-based models were approximately equivalent to the Context model (74% accuracy) for correct vs. incorrect discrimination. This suggests that for fine-grained detection of learner errors, a knowledge-based approach of SE content is more appropriate [19-21].

4 Conclusion

Several ITSs have incorporated the assessment of learner natural language responses using NLU techniques such as LSA, word-overlap, and other linguistic features. We tested which response features (content, context, combination) were most effective at accurately assessing SE quality, both in terms of correct vs. incorrect discriminations and classifying different error types. We were able to achieve moderate success at SE classification with models that included either the response content or response context, but there were no improvements when the models were combined.

Previous work on the classification of learner contributions has focused on response content [16-17, 22]. We expanded these previous efforts by also incorporating features of the context. We found that the effectiveness of content- and context-based features differed depending on the discrimination task. More specifically, the context-based model was sufficient to make correct vs. incorrect discriminations but the content-based model was needed for more specific error type classification. An effective approach for classification systems, then, would be to initially use context-based features to determine whether an SE is correct or incorrect. If the SE is classified as incorrect, the content features can then be used to make a finer-grain distinction between types of erroneous responses.

One interesting and informative finding was that we were relatively successful at making a general correct vs. incorrect SE classification *without even considering* the actual SE response. The success of this context model, which incorporated the learner's prior performance and other informative parameters, suggests that it can be used to make predictive assessments of SE quality. This information can be used to decide the optimal time to ask learners to provide an SE. However, this conclusion should be taken with a modicum of caution because further empirical testing of this

classification scheme will be necessary to determine how frequently SEs are misclassified and the impact this misclassification has on learning.

Automatic classification of SE quality and error-ridden reasoning has important implications for building adaptive and effective ITSs. Through the use of readily available context features as well as word-overlap comparisons, ITSs can use SEs to create a more accurate model of learner knowledge. ITSs can then use this information to provide individually tailored scaffolding based on errors identified in learner-generated explanations. This type of adaptive scaffolding will allow ITSs to more efficiently and effectively help learners to reach deeper levels of understanding.

Acknowledgement. The research was supported by the National Science Foundation (REC 0106965, ITR 0325428, HCC 0834847, DRL 1108845) and the Institute of Education Sciences (R305A080594). The opinions expressed are those of the authors and do not represent views of the NSF and IES.

References

1. Simon, H.: Problem solving and education. In Tuma, D. & Reif, F. (Eds.) Problem solving and education: Issues in teaching and research. Erlbaum, Hillsdale (1979)
2. Graesser, A., Jeon, M., & Dufty, D.: Agent technologies designed to facilitate interactive knowledge construction. Discourse Processes, 45(4), 298--322 (2008)
3. Prosser, M., & Trigwell, K.: Understanding learning and teaching. The Society for Research into Higher Education and Open University Press, Buckingham (1999)
4. Chi, M.: Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In Glaser, R. (Ed.) Advances in instructional psychology: Educational design and cognitive science (pp. 161--238). Erlbaum, Mahwah (2000)
5. Chi, M., Bassok, M., Lewis, M., Reimann, P., & Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145--182 (1989)
6. Chi, M., de Leeuw, N., Chiu, M., & LaVancher, C.: Eliciting self-explanations improves understanding. Cognitive Science, 18, 439--477 (1994)
7. Renkl, A., Stark, R., Gruber, H., & Mandl, H.: Learning from worked-out examples: The effects of example variability and elicited self-explanations. Contemporary Educational Psychology, 23, 90--108 (1998)
8. McNamara, D.: SERT: Self-explanation reading training. Discourse Processes, 38, 1--30 (2004)
9. Renkl, A.: Learning from worked-out examples: A study on individual differences. Cognitive Science, 21, 1--29 (1997)
10. Aleven, V., & Koedinger, K.: An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. Cognitive Science, 26, 147--179 (2002)
11. Conati, C., & VanLehn, K.: Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. International Journal of Artificial Intelligence in Education, 11, 398--415 (2000)
12. McNamara, D., Levinstein, I., & Boonthum, C.: iSTART: Interactive strategy trainer for active reading and thinking. Behavioral Research Methods, Instruments, and Computers, 36, 222--233 (2004)
13. O'Reilly, T., Best, R., & McNamara, D.: Self-explanation reading training: Effects for low-knowledge readers. In Forbus, K., Gentner, D., Regier, T. (Eds.) Proceedings of the

26th Annual Meeting of the Cognitive Science Society (pp. 1053--1058). Erlbaum, Mahwah (2004)
14. Williams, C., & D'Mello, S.: Predicting student knowledge levels from domain-independent function and content words In Kay, J. & Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems (pp. 494--503). Springer, Berlin (2010)
15. Pennebaker, J., Francis, M., & Booth, R.: Linguistic Inquiry and Word Count (LIWC). Erlbaum, Mahwah (2001)
16. Litman, D., Moore, J., Dzikovska, M., & Farrow, E.: Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In Dimitrova, V., Mizoguchi, R., DuBoulay, B., & Graesser, A. (Eds.),*P*roceedings of 14th International Conference on Artificial intelligence in education (pp. 149--156). IOS Press, Amsterdam (2009)
17. Graesser, A., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X.: Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.) Handbook of latent semantic analysis (pp. 243--262). Lawrence Erlbaum, Mahwah (2007)
18. Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.): The handbook of latent semantic analysis. Erlbaum, Mahwah (2007)
19. Aleven, V., Popescu, O., & Koedinger, K.: Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In Moore, J., Redfield, C., & Johnson, W. (Eds.) Proceedings of the 10th International Conference on Artificial intelligence in education (pp. 246--255). IOS Press, Amsterdam (2001)
20. Popescu, O., & Koedinger, K.: Towards understanding geometry explanations. In Rose, C. & Freedman, R. (Eds.) Building Dialogue Systems for Tutorial Applications, Papers of the 2000 AAAI Fall Symposium (pp. 80--86). AAAI Press, Menlo Park (2000)
21. Aleven, V., Popescu, O., & Koedinger, K.: Pilot-testing a tutorial dialogue system that supports self-explanation. In Cerri, S., Gouardères, F., & Paraguaçu, F. (Eds.) Proceedings of 6th International Conference on Intelligent Tutoring Systems (pp. 344--354). Springer, Berlin (2002)
22. Rus, V., McCarthy, P., Lintean, M., Graesser, A., & McNamara, D.: Assessing student self-explanations in an intelligent tutoring system. In McNamara, D. & Trafton, J. (Eds.) Proceedings of the 29th Annual Conference of the Cognitive Science Society (pp. 623--628). Erlbaum, Mahwah (2007)
23. Rus, V., McCarthy, P., & Graesser, A.: Analysis of a textual entailer. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico (2006)
24. Lehman, B., D'Mello, S., Strain, A., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A.: Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In Biswas, G., Bull, S., Kay, J., & Mitrovic, A. (Eds.) Proceedings of 15th International Conference on Artificial Intelligence in Education (pp. 171--178). Springer-Verlag, Berlin (2011)
25. Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A.: Utterance classification in AutoTutor. In Burstein, J., & Leacock, C. (Eds.) Building Educational Applications using Natural Language Processing: Proceedings of the HLT - NAACL Conference 2003 Workshop (pp. 1--8). Association for Computational Linguistics, Philadelphia (2003)
26. Baayen, R., Piepenbrock, R., & Gulikers, L.: The CELEX lexical database (Release 2) [CD-ROM]. University of Pennsylvania, Linguistic Data Consortium, Philadelphia (1995)
27. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1), 10 (2009)