Assessing Question Quality Using Natural Language Processing

Authors

Affiliations

Abstract. iSTART (Interactive Strategy Training for Active Reading and Thinking) is an intelligent tutoring system (ITS) designed to improve reading comprehension by providing strategy instruction. The goal of this study is to assess the potential to integrate a question asking module into iSTART that will provide automated feedback to students' based on their generated questions. Such a system will require a natural language processing (NLP) algorithm to assess readers' questions and provide feedback. In this study, we take the initial steps toward developing and testing a NLP algorithm using a corpus of 4629 questions collected from Amazon's Mechanical Turk participants. Participants read short texts and generated questions for each text read. These questions were coded by human raters using a four-level taxonomy that classified questions as shallow to deep. NLP indices were calculated for each generated question and machine learning techniques were used to predict the human ratings. The findings indicate that indices related to lexical sophistication were able to modestly predict question classification. Predictions improved when NLP indices were used to predict two categories (shallow vs. deep).

Keywords: Intelligent Tutoring Systems, Artificial Intelligence, Natural Language Processing, Educational Technology Design, Question Classification

1 Introduction

Students often struggle to comprehend difficult texts [1]. Although the use of active reading strategies can help improve comprehension, students rarely use them while reading [2]. Instruction and extended practice can promote the use of such reading strategies [3]. Unfortunately, educators limited by time constraints are typically unable to provide immediate feedback to individuals on students' attempts to apply these strategies. Thus, Intelligent Tutoring Systems (ITSs) that can provide instruction and automated feedback on reading strategies are uniquely poised to meet students' needs. iSTART (Interactive Strategy Training for Active Reading and Thinking) is an ITS that provides instruction on self-explanation strategies (i.e., comprehension monitoring, paraphrasing, prediction, bridging, and elaboration) and generative strategy practice. Generative practice modules provide opportunities for extended practice of reading

strategies with immediate feedback using natural language processing (NLP; [4]). Research indicates that iSTART improves learners' ability to construct quality self-explanations and increases reading comprehension [5,6].

Similar to self-explanation, question asking is an effective reading strategy [7], and asking deep (i.e., questions that get at a deeper form of knowledge) rather than shallow questions (i.e., questions that address isolated facts) during reading improves reading comprehension. Researchers have created systems to *generate* questions for learners to answer during learning [8,9]. However, to our knowledge, no systems are available to *assess* the quality of questions that readers ask *during* reading. Thus, an overarching goal of this project is to incorporate into iSTART a module that can provide automated feedback on students' generated questions. In such a module, students would practice asking deep questions (while reading texts) and receive automated immediate feedback through NLP algorithms that assess question quality (i.e., deep vs. shallow). The goal of the current investigation is to identify NLP indices predictive of question quality and assess the accuracy of predicted scores compared to human ratings.

2 Theoretical Background

The purpose of a question is to fill a gap or deficit in knowledge. A question is a specific speech act that acts as an inquiry, typically followed by a question mark. They are different from other speech acts (e.g., statements) in that the intention is to seek information [10]. Asking questions during reading promotes active learning and comprehension monitoring [3]. Rosenshine, Meister, and Chapman [7] reported a review of question instruction intervention studies; the review revealed improved comprehension with an effect size of .36 on standardized assessments and .86 on experimenter generated materials. Furthermore, research has generally found that "deeper" question asking improves comprehension compared to "shallow" [11,12].

What distinguishes a deep from a shallow question? It is well established that researchers consider knowledge to exist on varying levels from shallow to deep. Perhaps the most commonly referenced, Bloom's Taxonomy [13], identifies the cognitive processes associated with different forms of knowledge acquisition. On the shallow end of this scale are memory search processes such as recognition and recall. On the deeper end are reasoning processes such as synthesis and evaluation. Different types of questions promote different types of processes and require differing levels of knowledge to answer [14]. For example, a shallow question may require a simple yes or no answer (e.g., Is it sunny outside today?). A deeper question may ask to describe a procedure (e.g., What steps would one take to ensure financial security?).

Multiple fields of research (i.e., psychology, education, computational linguistics) have developed question taxonomies. For example, from the field of education, Mosenthal [15] proposed a five-point scale categorizing questions on concreteness (1 = most concrete) to abstractness (5 = most abstract). Similarly, Goldman and Duran [16] assessed question asking during reading and identified five classes of questions and how they relate to the text. According to this scale, shallow questions require surface-level

processing of information and share a verbatim representation of the text. Deep questions require reasoning beyond what was in the text.

There are available ITSs including question-generation mechanisms to provide questions that promote active learning of deep conceptual material. For example, AutoTutor is an ITS that has been used to successfully teach topics in conceptual physics [8,9] and computer literacy [17]. It simulates a human tutor by holding a conversation with the learner using natural language. Rather than building a *question asking or generation mechanism* such as the one in systems like AutoTutor, our goal is to create a mechanism that *provides feedback on the questions* students ask while reading. In pursuit of this goal, the first step is to create an algorithm that can distinguish between deep vs. shallow level questions.

3 Categorizing Questions

According to Graesser and Person [10], students rarely ask questions during classroom interactions, and the questions that are generated are frequently 'unsophisticated'. The current investigation used a question coding scheme based on the Graesser and Person [10] question taxonomy. Their question taxonomy organizes 16 types of questions into two broad categories, short answer and long answer. Short answer questions require the answerer to provide a word or phrase as a response. Long answer questions normally require several sentences to answer. The authors further distinguish certain questions as deep-reasoning questions, which require "reasoning in logical, causal, or goal oriented systems" [10; p. 112]. One such deep-reasoning question type, *causal consequence* questions ask "What are the consequences of an event or state?" An example of a causal consequence question is "What happens if you call 911 in a non-emergency?"

To our knowledge there are surprisingly few instances where researchers attempted to automatically assess the types of questions learners ask. Olney et al. [18] used a shallow approach to classify speech utterances during tutoring AutoTutor sessions. Other researchers have adopted their utterance classification approach to assess student interaction within ITSs [19,20] and within the classroom [21]. The AutoTutor algorithm categorized questions based on question stems (i.e., who, what, how), question mark punctuation, and keyword matching. Olney et al. [18] showed that they could successfully classify 11 of the 16 categories from the Graesser and Person [10] taxonomy. However, their corpus contained relatively few questions compared to all utterances recorded during tutoring sessions (3% of over 9000 utterances were questions). By contrast, in our context, readers were explicitly instructed to ask questions (as opposed to any utterance), resulting in over 4000 questions. Human coders then applied a classification scheme modified from Graesser and Person [10] to classify the questions, thus producing the data for the development of the NLP algorithm described in this study.

4 Method

4.1 Participants

Two hundred thirty-three participants were recruited using the Amazon Mechanical Turk (MTurk) online research service. Ninety-one participants did not complete the question generation task; responses from participants who did not complete the task were included in the question corpus. In MTurk, voluntary workers receive financial compensation in exchange for completing Human Intelligence Tasks (HITs). Compared to the university subject pool used in most behavioral research, workers in MTurk represent a diverse population, and the collection of data over the Internet eliminates the potential for unintended experimenter effects [22]. No demographic data was collected.

4.2 Materials

The 30 texts used for the corpus collection were obtained from the California Distance Learning Project (CDLP; www.cdlponline.org), with the permission of the Sacramento County Office of Education. Texts on the CDLP website are simplified news articles that are used to help improve the reading skills of adult literacy. The texts are life-relevant to adult learners, and include topics such as health and safety, housing, family, and money. Each text was between four and seven paragraphs, and ranged in word count from 128 to 452 words (SD = 52.7 words). Flesch-Kincaid grade level was between 3rd and 8th grade (SD = 1.0) for all texts. Each participant read three short texts, randomly selected from the full set of 30 texts. Before data collection began, a researcher identified between three and seven target sentences (total of 164 target sentences) in each text where participants were required to ask questions. The length and complexity of the texts and sentences were used to select target sentences. Sentences that could be expanded on using existing knowledge or previous text content were selected as targets.

4.3 Procedure

Qualtrics (www.qualtrics.com) survey software randomly selected and presented the texts to participants. Texts were displayed in chunks, delimited by target sentences. For example, if the first target sentence was sentence four, four sentences were initially shown on the screen, with the target sentence displayed in bold. Below each text chunk was a prompt to write a question for the target sentence. Participants were allotted up to 6 minutes to write a question before the survey software automatically submitted their response. They were allowed to submit a response after 1 minute had elapsed. Once the response was submitted, the participant moved to the next sentence. The survey software recorded participant responses and the time taken to submit each response.

4.4 Corpus

Average response time to generate each question was 85.2 seconds (SD = 33.9). The initial dataset included 4,629 responses, ranging from 1 to 58 words. Because the survey software randomly selected texts, responses were generated by between 16 and 30 participants for each text; the dataset included between 81 and 212 responses for each text. When a response included multiple questions (e.g., "How will it get fixed? When will it get fixed?"), they were separated; this resulted in a set including between 20 and 40 responses per target sentence. Additionally, some responses were not used in our analyses because they were statements, not questions (e.g., "Flight attendants need to learn how to protect themselves in a bad situation."). After this initial data cleaning, the final dataset included 4,575 questions that were coded by two trained researchers.

4.5 Human Question Coding

A question coding scheme was developed to code each question. The scheme was based on the Graesser and Person [10] question taxonomy. We modified the coding scheme slightly, based on initial training of the coding scheme. The original taxonomy distinguishes between 16 different question types (e.g., verification, definition, etc.). Two question types (instrumental/procedural and enablement) were collapsed into one type. Initial rounds of coding subsets of the data indicated the coders were unable to reliably distinguish between those two types. We then organized the question types into four categories ranging from (1) very shallow to (4) very deep. Shallow level 1 requires a one-word answer (e.g., yes/no, a value, a name of a person). Shallow level 2 still requires very short responses but may include more than one element (e.g., providing a definition). Deep level 3 requires longer responses but do not address causal mechanisms of a system (e.g., comparing one entity to another). Deep level 4 questions require lengthy responses that address causal mechanism underlying system functioning (e.g. determining what antecedent led to an outcome).

Two trained researchers coded the data. They first went through two initial rounds of applying the coding scheme a subset of approximately 20% of the question corpus and discussing discrepancies. Interrater reliability for each of these training rounds was established using Cohen's kappa, bivariate correlation (r), percent exact agreement, and percent adjacent agreement. The kappa scores reported were calculated using linear weighting. Adjacent agreement was calculated using the number of cases in which the two coders had either exact (e.g., A rating = 2 & B rating = 2) or the difference between the levels of codes is not more than one (e.g., A rating = 2 & B rating = 3). Interrater reliability improved from the first round of training (kappa = .74; r = .78; 76% exact agreement; 92% adjacent agreement) to the second round (kappa = .80; r = .86; 79% exact agreement; 95% adjacent agreement).

The coders next coded 60% of the data set each, with 20% overlap to establish the final interrater reliability. Agreement on the final subset used for reliability was: kappa = .84, r = .67, 82% exact agreement, and 92% adjacent agreement. Remaining differences between the codes from the two researchers were resolved with discussions; these discussions resulted in a finalized score for each question, from one to four.

4.6 Linguistic Variables

The generated questions were separated and cleaned to remove any non-linguistic data. Each question was run through a number of NLP tools including the Tool for the Automatic Analysis of Lexical Sophistication (TAALES), the Tool for the Automatic Analysis of Cohesion (TAACO) and the Constructed Response Analysis Tool (CRAT). The selected tools reported on language features related to lexical sophistication, text cohesion, and overlap between the text and the questions generated respectively. The tools are discussed in greater detail below.

TAALES. TAALES [23] is a computational tool that is freely available, easy to use, works on most operating systems (Windows, Mac, Linux), allows for batch processing of text files, and incorporates over 150 classic and recently developed indices of lexical sophistication. These indices measure word frequency, bi-gram and tri-gram frequency, and range counts taken from a number of corpora including the Corpus of Contemporary American English (COCA) [24], academic words and phrases, word information measures reported by the MRC Psycholinguistic Database [25], and word association metrics calculated using COCA.

TAACO. TAACO [26] incorporates over 150 classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates a part of speech (POS) tagger and synonym sets from the WordNet lexical database [27]. TAACO provides linguistic counts for POS tags for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type-token ratio (TTR) indices (for all words, content words, function words, and n-grams), sentence overlap indices that assess local cohesion for all words, content words, function words, POS tags, and synonyms, paragraph overlap indices that assess global cohesion for all words, content words, function words, POS tags, and synonyms, and a variety of connective indices including opposition connectives (e.g., *but, however, nevertheless*).

CRAT. CRAT [28] is an engine that calculates indices related to a) the linguistic and semantic similarities between a source text and a constructed response (i.e., the generated question), b) the linguistic sophistication of a constructed response, and c) text properties (e.g., length and syntactic categories). The similarity indices include lexical similarity calculated using key word overlap, synonym overlap, and latent semantic analysis (LSA) similarity [29] and phrasal similarity calculated using key bigram and trigram overlap and key part of speech sensitive slot-grams (e.g., a trigram with an open slot such as *into the* ____).

4.7 Statistical Analysis

We used the indices reported by the NLP tools to predict human scores (1 through 4) for the corpus of questions. Indices reported by the tools that lacked normal distributions were removed. A multivariate analysis of variance (MANOVA) was conducted to examine which indices reported differences between the four question categories.

The MANOVA was followed by stepwise discriminant function analysis (DFA) using selected NLP indices that demonstrated significant differences among the four categories of questions but did not exhibit multicollinearity ($r > .90$) with other indices in the set. In the case of multicollinearity, the index demonstrating the largest effect size was retained in the analysis. The DFA provides an algorithm to predict group membership (i.e., whether the question was scored a 1, 2, 3, or 4) through a discriminant function coefficient. A DFA model was first developed for the entire corpus and this model was then used to predict group membership of the questions using leave-one-out-cross-validation (LOOCV) in order to ensure that the model was stable across the dataset. A follow up analysis was conducted in which the questions were separated into binary categories. In this case, questions scored as 1 or 2 were categorized as shallow and questions scored as 3 or 4 were categorized as deep. The same statistical procedure was followed for the second analysis.

5 Results

5.1 Four Category Analysis

A MANOVA was conducted using the NLP indices as the dependent variables and the four categories of questions as the independent variables. Fifty-two variables demonstrated significant differences between the question scores. These 52 variables were used in the DFA. The DFA retained 28 variables. The majority of these variables were related to lexical sophistication. One variable related to semantic similarity between the text and the question was retained as were two variables related to type-token ratio counts (see Table 1 for the MANOVA results for the variables retained in the DFA). The results demonstrate that the DFA using these 28 indices correctly allocated 1904 of the 4575 questions in the total set, $\chi 2$ (df=9, n=4575) = 669.567, $p < .001$, for an accuracy of 41.6%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis also allocated 1834 of the 4575 texts for an accuracy of 40.1%. The measure of agreement between the human produced question score and that assigned by the model produced a weighted Cohen's Kappa of 0.21.

Table 1. List of Indices and MANOVA Results for Four Category Analysis

Index	Greater at deeper level±	*F*	Partial N^2
Proportion of bigrams COCA (70,000 words)	Yes	39.247**	0.025
Average lexical decision accuracy	Yes	30.276**	0.019
Lemma TTR (content words)	Yes	24.608**	0.016
Log content word range COCA news	Yes	18.724**	0.012
Lemma overlap between question and text	Yes	18.945**	0.012
Mean combined concreteness score	Yes	18.634**	0.012
Word frequency: Thorndike Lorge (all words)	No	15.015**	0.010
Word frequency (log): BNC spoken content words	Yes	14.082**	0.009
Word frequency (log): COCA spoken content words	Yes	10.377**	0.007
Proportion of bigrams COCA (80,000 words)	No	10.964**	0.007
Lemma TTR (news words)	Yes	6.965**	0.005
Proportion of bigrams COCA (50,000 words)	Yes	7.105**	0.005
Mean COCA bigram log frequency score	Yes	8.030**	0.005
Lemma TTR (COCA fiction)	Yes	7.967**	0.005
Standardized naming RT	No	5.911**	0.004
Bigram proportion score COCA (100,000 words)	Yes	6.374**	0.004
Lemmas TTR (magazine words)	Yes	5.944**	0.004
Semantic variability of contexts	Yes	6.352**	0.004
Lemma TTR (academic words)	No	3.949*	0.003
Lemma TTR (all words)	Yes	5.016*	0.003
Bigram proportion score BNC written words	Yes	4.044*	0.003
TTR for questions (content words)	Yes	4.098*	0.003
Academic bigram association strength (COCA)	Yes	5.085*	0.003
Bigram proportion score COCA (60,000 words)	No	3.436*	0.002
Lemma proportion COCA (fiction)	Yes	2.477*	0.002
Word frequency: COCA academic function words	No	3.094*	0.002
Word frequency: COCA spoken content words	Yes	2.967*	0.002
Log academic word range COCA (all words)	No	2.772*	0.002

* $p < .05$, ** $p < .01$; TTR = type-token ratio

± Yes indicates average value for deep questions (level 3 and 4) was above the overall mean

5.2 Two Category Analysis

A MANOVA was next conducted using the NLP indices as the dependent variables and the two categories of questions as the independent variables. Twenty-five variables demonstrated significant differences between the two question scores. These 25 variables were used in the DFA. The DFA retained 14 variables. The majority of these variables were related to lexical sophistication with two variables related to type-token ratio counts retained (see Table 2 for the MANOVA results for the variables retained in the DFA). The results demonstrate that the DFA using these 15 indices correctly allocated 2817 of the 4575 questions in the total set, $\chi 2$ (df=1, n=4575) = 245.063, p < .001, for an accuracy of 61.6%. For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 2794 of the 4575 texts for an accuracy of 61.1%. The measure of agreement between the human produced question score and that assigned by the model produced a weighted Cohen's Kappa of 0.23.

Table 2. List of Indices and MANOVA Results for Two Category Analysis

Index	Greater at deeper level±	*F*	Partial N^2
Average lexical decision accuracy	Yes	86.186**	0.018
Mean combined concreteness score	Yes	38.952**	0.008
Word frequency (log): BNC spoken (all words)	Yes	37.730**	0.008
Word frequency: Thorndike Lorge (all words)	No	31.145**	0.007
Word frequency (log): COCA spoken content words	Yes	24.156**	0.005
Word range COCA news (content words)	Yes	23.446**	0.005
Content words TTR	Yes	21.350**	0.005
Semantic similarity across words in question	Yes	14.107**	0.003
Standardized naming reaction time across all participants for this word	No	14.244**	0.003
Word frequency (log): BNC (all words)	No	9.924*	0.002
Lemma TTR	Yes	7.480*	0.002
Lemma proportion COCA	No	7.397*	0.002
Bigram proportion score BNC written words	Yes	5.767*	0.001
Bigram proportion score COCA (60,000 words)	No	4.911*	0.001

* p < .05, ** p < .01; TTR = type-token ratio

± Yes indicates average value for deep questions (level 3 and 4) was above the overall mean

6 Discussion and Conclusion

Although asking deep questions while reading can improve comprehension, learners rarely ask questions during learning. An overarching objective of this project is to create a module within iSTART which would promote learners to ask deep questions during

reading with the goal of improving comprehension. For example, deep questions would require learners to identify a causal mechanism. A shallow question usually requires a one-word answer such as asking for verification (e.g., Is it the case that X is true?). We found that when considering four levels of question quality (shallow to deep), NLP indices could modestly predict question category (based on human coding). When considering only two levels (shallow and deep), the predictive abilities of the NLP indices improved. Results revealed that the most predictive indices relate to lexical sophistication and lexical and semantic overlap.

Specifically, the results indicated that deeper level questions contained less sophisticated words and greater lexical and semantic overlap both within the question and with the text. In terms of lexical sophistication, deeper level questions included words with higher accuracies on lexical decision tests, more frequent words, less specific words, and more concrete words. These results indicate that deeper level questions contain words that are not more sophisticated, but rather use words that are easier to process and more familiar allowing for better comprehension of the question.

Deeper level questions also are more cohesive, which should also afford better comprehension. In terms of cohesion, deeper level questions repeat words more often (i.e., report a higher TTR), reported greater semantic similarity within the question (i.e., high LSA values across words in the question), and have greater overlap with the text from which they are derived. In total, these linguistic features indicate that deeper level questions are easier to process in terms of language complexity.

While speech act classifications have been successfully incorporated into systems to drive interactional moves between the learner and the system [30], there are surprisingly few instances where researchers have attempted to automatically assess the types of questions learners ask. As opposed to the syntactic and semantic features of the questions used in our investigation, Olney et al. [18] used surface features of the questions (e.g., frozen expressions and punctuation). Their question corpus also included fewer than 300 questions, representing only 3% percent of the total utterances. Hence, their findings have limited applicability in our context where the students are being explicitly prompted to generate questions. Nonetheless, our approach may benefit from combining the use of surface features and syntactic and semantic properties in the future.

Our investigation is not without limitations. One challenge to any classification project is to use the appropriate classification scheme. We chose to use an existing taxonomy developed to identify the type of knowledge that questions would elicit. Given that the Graesser and Person [10] taxonomy was developed for identifying the types of questions asked during tutoring; it is possible that an alternative question taxonomy would be more appropriate in this context. Another challenge is the selection of NLP indices to categorize questions. The indices used for this study were originally designed to assess constructed responses to open response questions [28] and language development in terms of discourse approaches (i.e., cohesion) and lexical sophistication. Our results may be improved by an analysis of surface features (e.g., question stem identification, phrase identification, keyword matching), combined with NLP indices developed to identify questions and similar discourse markers.

Nonetheless, the current study makes a significant contribution to the literature by taking strides towards automating classifications of question quality. This study also

contributes to the improvement of an existing ITS with the objective of enhancing reading comprehension for a wide range of readers (iSTART:[5,6]). Our hope is that future work that builds on this foundation will be beneficial to the development of other ITSs and a variety of computer-based learning environments.

7 References

1. Snow, C.: Reading for understanding: Toward an R&D program in reading comprehension. Rand Corporation (2002)
2. Pressley, M., & Ghatala, E. S.: Self-regulated learning: Monitoring learning from text. Educational Psychologist, 25(1), 19-33 (1990)
3. Palincsar, A. S., & Brown, A. L.: Reciprocal teaching: Activities to promote read (ing) with your mind. Reading, thinking and concept development: Strategies for the classroom. New York: The College Board. (1985)
4. McNamara, D. S.: Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. Discourse Processes (2015)
5. Jackson, G. T., & McNamara, D. S.: Motivation and performance in a game-based intelligent tutoring system. Journal of Educational Psychology, 105(4), 1036 (2013)
6. McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y.: Improving adolescent learners' reading comprehension with iSTART. Journal of Educational Computing Research, 34(2), 147-171(2006)
7. Rosenshine, B., Meister, C., & Chapman, S.: Teaching students to generate questions: A review of the intervention studies. Review of educational research, 66(2), 181-221 (1996)
8. Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., ... & Person, N. K. : Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society (pp. 1-6) (2003)
9. VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P.:When are tutorial dialogues more effective than reading? Cognitive Science, 31, 3-62 (2007)
10. Graesser, A. C., & Person, N. K.: Question asking during tutoring. American educational research journal, 31(1), 104-137 (1994)
11. Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L.: Impact of question-answering tasks on search processes and reading comprehension. Learning and Instruction, 19(1), 13-27 (2009)
12. Vidal-Abarca, E., & Sanjose, V.: Levels of comprehension of scientific prose: The role of text variables. Learning and Instruction, 8(3), 215-233 (1998)
13. Bloom, B. S.: Taxonomy of educational objectives. Vol. 1: Cognitive domain. New York: McKay, 20-24 (1956)
14. Graesser, A. C., & Franklin, S. P.: QUEST: A cognitive model of question answering. Discourse processes, 13(3), 279-303 (1990)
15. Mosenthal, P. B.: Understanding the strategies of document literacy and their conditions of use. Journal of Educational Psychology, 88(2), 314 (1996)
16. Goldman, S. R., & Durán, R. P.: Answering questions from oceanography texts: Learner, task, and text characteristics. Discourse Processes, 11(4), 373-412 (1988)
17. Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., & Person, N.: AutoTutor improves deep learning of computer literacy: is it the dialog or the talking head? In U. Hoppe, F. Verdejo, & J. Kay (Eds.), Proceedings of artificial intelligence in education (pp. 47e54) Amsterdam: IOS Press (2003)

18. Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A.: Utterance classification in AutoTutor. In Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2 (pp. 1-8). Association for Computational Linguistics (2003)
19. Li, H., Samei, B., Olney, A. M., Graesser, A. C., & Shaffer, D. W.: Question classification in an epistemic game. In 3rd Workshop on Intelligent Support for Learning in Groups (ISLG) at the 12th International Conference on Intelligent Tutoring Systems, Springer (2014)
20. Soh, L. K., Khandaker, N., Liu, X., & Jiang, H.: A computer-supported cooperative learning system with multiagent intelligence. In Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems (pp. 1556-1563). ACM (2006)
21. Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., ... & Graesser, A.: Domain Independent Assessment of Dialogic Properties of Classroom Discourse. Grantee Submission (2014)
22. Crump, M. J., McDonnell, J. V., & Gureckis, T. M.: Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one, 8(3), e57410 (2013)
23. Kyle, K., & Crossley, S. A.: Automatically assessing lexical sophistication: Indices, tools, findings, and application. Tesol Quarterly, 49(4), 757-786 (2015)
24. Davies, Mark.: The corpus of contemporary American English. BYE: Brigham Young University (2008)
25. Coltheart, M.: The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology Section A, 33, 497–505 (1981)
26. Crossley, S. A., Kyle, K., & McNamara, D. S.: The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods, 48(4), 1227-1237 (2016)
27. Miller, G. A.: WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39–41 (1995)
28. Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S.: Automatic assessment of constructed response data in a chemistry tutor. In T. Barnes, M. Chi, & M. Feng (Eds.), Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016), (pp.336-340). Raleigh, NC: International Educational Data Mining Society (2016)
29. Landauer, T. K., Foltz, P. W., & Laham, D.: An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284 (1998)
30. Samei, B., Li, H., Keshtkar, F., Rus, V., & Graesser, A. C.: Context-based speech act classification in intelligent tutoring systems. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), Proceedings of the International Conference on Intelligent Tutoring Systems, (pp. 236-241). New York: Springer (2014)