Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America
{mheilman,kct,callan,max}@cs.cmu.edu

Abstract

We present an intelligent tutoring system called REAP that provides reader-specific lexical practice for improved reading comprehension. REAP offers individualized practice to students by presenting authentic and appropriate reading materials selected automatically from the web. We encountered a number of challenges that must be met in order for the system to be effective in a classroom setting. These include general challenges for a system that uses authentic materials, as well as more specific challenges that arise from integrating the system with pre-existing classroom curricula. We discuss how these challenges were met, and present evidence that REAP has gained acceptance into the classroom at the English Language Institute at the University of Pittsburgh.

1. System Description

We begin with brief descriptions of the REAP intelligent tutoring system and its primary users. For a more detailed description of the REAP project, please see [1] and [2]. The REAP project's goal is to provide appropriate, authentic reading materials to students learning to read. It gathers and selects documents automatically from the web, which raises a number of concerns that will be discussed in this paper. The system has focused on English language so far, but future developments could extend the scope of the project to other languages. REAP incorporates a variety of statistical language modeling and information retrieval methods in order to model students' knowledge and find useful reading passages for them.

Recent work on the REAP system includes creating a system for the University of Pittsburgh's English Language Institute (ELI) Reading 4 course, an upper-level course for English as a Second Language (ESL) that focuses on reading skills. A study on usability of REAP is currently in progress at the ELI. In this study, which we will refer to as the Spring '06 ELI Study, thirty-three students use the system once a week for forty minutes over the course of the semester, reading documents containing target unknown vocabulary identified from a pre-test.

REAP gathers documents from the Web in order to find useful, authentic reading material for these students. The documents are analyzed according to syntactic features, readability, length, and the occurrence of target vocabulary. The tutor uses an extended version of the Lemur Toolkit for Language Modeling and Information Retrieval [3] to annotate the documents and create an index for language-model based retrieval. When a student uses REAP, the system searches among this set of documents for those that satisfy a number of constraints, including the student's target vocabulary list, document length, his or her user model, and the target reading level for the course, which is sixth to eighth grade. After reading a document, usually from one to two pages in length, the student works through a series of automatically generated exercises based on the target vocabulary found in the reading. The student model is updated after every reading so that the optimal document can be retrieved for the next reading passage.

By using authentic reading materials the REAP system offers realistic training and individualized curricula to students. Reading textbooks and hand-selected materials are usually well-controlled, appropriate, and contain high-quality input, yet such materials are also static, difficult to produce, and very limited in quantity. In a classroom setting, it is typical that all students see the same material from a textbook, regardless of the state of their lexical or grammatical development. Also, reading materials for use in most classrooms must meet a wide variety of syntactic and lexical constraints in order for students of a given reading proficiency to be able to read them without confusion. Teachers or textbook authors often have to heavily edit or even produce the reading materials themselves in order to meet these constraints, introducing some amount of artificiality into the materials. Intelligent tutoring systems such as REAP can examine large corpora such as the Web in order to automatically select materials that meet these various criteria. Students using REAP work toward their ultimate goal of reading real text by actually reading real text.

Intelligent tutoring systems also provide students with individualized practice rather than static sets of exercises. Students go through readings at very different rates, and so faster students need a greater number of more difficult reading passages than do slower students. In a current study ten students using the REAP system had completed fewer than ten reading passages, while twelve students had completed twenty or more, despite having the same time on task. The average number completed was just under seventeen. REAP selects as many documents as are necessary for each student, and these documents satisfy certain lexical, syntactic, and readability constraints based on a model of the current student's knowledge. Finding a large number of appropriate documents for an entire classroom of students can in many cases only be accomplished by an intelligent tutoring system. Such a system is therefore very valuable to language teachers.

The value of the system is demonstrated in results from an exit survey taken toward the end of a recent study, shown in Figure 1. The students (N=33) were asked to respond on a Likert scale from 1 to 5 indicating the degree to which they

agree to given statements about the system. The results indicate that students feel that the REAP system is easy to use, valuable for learning both target and non-target vocabulary, and worth using in future classes. Students wanted more personalization and choice of the reading topics, however, to make the system more engaging.

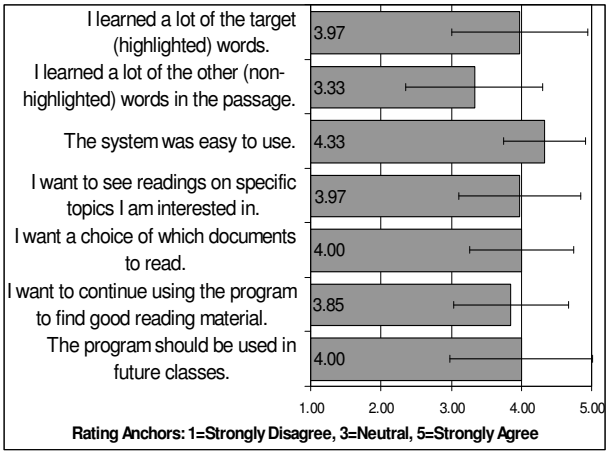

Figure 1 *Opinions about REAP from an exit survey.*

Preliminary results from the ELI Spring '06 study also indicate that students are learning their target words. After each reading, students work through automatically generated cloze exercises related to the target words from the passage. These exercises are discussed in more detail later in the paper. The average percentage of these exercises answered correctly during each session has increased over the course of the semester, though not to the level of statistical significance. This trend is shown in Figure 2, in which the percentage of exercises answered correctly is plotted against the time in days since the start of the study. The mean, minimum, and maximum values for the percentage of correctly answered exercises by a given student over the entire semester were 85.0%, 44.4%, and 98.8%, respectively. There were a total of 2339 exercises following 902 reading passages for all students at all dates.

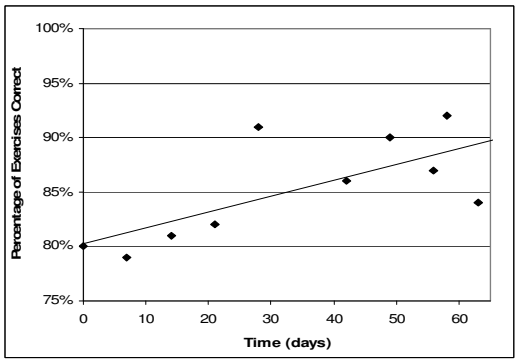

Figure 2 *Percentage Correct on Post-reading exercises over time*

2. Challenges for REAP

There is a large set of criteria and specifications that have to be satisfied in order for the REAP system to be effective in a classroom setting. The system has to be able to provide a large number and wide variety of documents to students. These documents also have to be filtered by the occurrence of target vocabulary, document length, reading level, text quality, topic, and perhaps even writing style. The system also has to present readings in an order and manner that is optimally effective and efficient for student practice. Finally, the system has to be capable of generating exercises that are relevant to the reading passages and that promote learning of the target vocabulary.

Finding and filtering documents is a significant challenge for the REAP system. While any search engine can find documents that contain a certain target vocabulary word, it is no simple task to find documents about appropriate topics that contain useful contextual information and consist of well-formed sentences at the appropriate reading level for language learners. In the Spring '06 ELI study, the target reading level was between sixth and eighth grade according to first language grade levels. The system employs a language modeling approach developed by Collins-Thompson and Callan [4] that creates a model of the lexicon for each grade level and predicts readability of given documents according to those models. For web documents, they found that language modeling-based prediction has a much stronger correlation with human-assigned levels than other readability measures. This automatic readability measure allows the tutoring system to select materials that are of the appropriate difficulty.

The system also filters documents by their syntactic and organizational quality in order to provide students with coherent reading passages. In early stages of the system, poorly organized documents were a major point that the ELI curriculum supervisor focused on. Web documents are in many cases not very well organized, but rather written in a very informal style that can be confusing to students. In early versions of REAP, students would occasionally see message board postings, advertisements and commercial sites, and even documents consisting solely of menus and links to other web pages. REAP uses parser confidence scores from the Stanford Parser [5] for the text in a document to calculate a text quality measure based on syntactic well-formedness. Besides the rare case, REAP presents only well-formed documents. Previously, the curriculum supervisor felt it would be necessary to review every single document before it would be presented to students; she now trusts the automatic filtering to perform that task.

Being accepted into the ELI classroom required that the system be compatible with the course curriculum. The ELI therefore wanted the system to present documents that contained a subset of the Academic Word List [6], which consists of words that are deemed important for incoming college undergraduates. These words are rare and fairly difficult for learners (e.g., "subsidiary," "reliance," "amendment"), and so do not appear very often in documents of the target grade level range, sixth to eighth grade. What is more, the system should ideally present documents that contain two or more of these words together in order to accelerate the student's progress through the curriculum. Many of the target words, however, are unrelated to each other (e.g., "transmission" and "sacred"),

making it very difficult to find useful documents containing more than one target word.

After filtering out documents that contain only a single target word, are of inappropriate reading level, are too long, or do not contain many well-formed sentences, only about 0.5% of the documents remain. For some words, there were fewer than five useful documents in a database of over 50,000 documents that contained at least one target word. In order to build this database of documents with target vocabulary, the system searched through millions more documents. It is therefore a significant challenge to find a sufficient number of documents that contain specific target words.

The ELI has brought up a number of other issues besides the readability and syntactic quality of passages that we have started to address. The topics and the contextual information of reading materials are primary concerns. Many documents such as legal proceedings, UNIX manual pages, and articles about local politics are uninteresting to their students. Also, many documents are news articles that are written specifically for an audience that is already familiar with the subject. In addition, some topics (e.g., terrorism, war) are sensitive to the international students at the ELI and should be avoided. What is more, there is often a mismatch between the reading level of second language students and their interests. While ESL students in college may have a sixth grade reading level, a large portion of the documents on the Web that are written at that reading level cover topics that are not interesting to adults. We have implemented some simplistic, topic-specific constraints that filter out documents if certain words occur, and are now trying to consistently provide material that is engaging to students. Another problem cited originally by the ELI is that many of the readings contained slang from other English-speaking countries. Most of the students at the ELI are learning English in order to work or attend school in America, and so Australian or British slang is unfamiliar and confusing to students. The REAP system has filtered out documents from non-U.S. domains in order to solve this problem. The ELI also originally requested that students only see narratives, or stories, since they are easier for beginning students to read. Although we would like to provide such a feature, automatically identifying the writing style of documents is very difficult, but will be addressed in the future.

In addition to the selection of reading materials, the proper presentation of these materials is also a significant concern for the users of the REAP system. For ease of use by students, REAP presents documents within a web browser-based application. The system strips outside links on the web pages and highlights target vocabulary words. The ELI also requested that students have access to a dictionary, so we implemented this feature by using a research-licensed version of the Cambridge Advanced Learners Dictionary [7]. The dictionary allows the students to access easy-to-grasp definitions for any unknown words that they encounter while reading. Incorporating an electronic dictionary into REAP also allows teachers and researchers to track dictionary use by the students, something that is not feasible when using a paper dictionary. As with any dictionary, multiple definitions are presented for each word, often for different parts of speech. We plan to incorporate part of speech tagging and word sense disambiguation so that the multiple definitions for each word can be ordered according to the context in which the word was used.

Intelligent tutoring systems must engage the student in active learning, so it was important to create appropriate exercises to follow the readings in the REAP tutor. Ideally, these exercises would be production tasks where the student is asked to use a given word in a sentence or even provide a definition of that word, but automatically and accurately assessing the correctness of student answers to such questions is extremely difficult. The system can, however, present a variety of multiple choice questions, as described by Brown et al. [8]. The system currently uses cloze questions, in which the student must select the most appropriate word to complete a sentence. These exercises are generated automatically from our corpus by choosing sentences that contains contextual clues that sufficiently narrow down the possible responses. For a previous study in the fall of 2005 using REAP, the ELI teachers chose to write the exercises by hand. Since that time the quality of the question generation tool has improved, however. For the ELI Spring '06 study, the questions were generated automatically and then filtered manually by teachers. We intend for future versions of the system to produce high quality cloze questions fully automatically.

3. Gradual Acceptance into the Classroom

Although a few of the original specifications have not yet been met, REAP has made a great deal of progress toward gaining acceptance into the classroom. We were able to resolve the major issues related to the quality, availability, and presentation of reading materials, and the ELI now sees the system as a valuable teaching tool. Next semester, the ELI will assign grades to students for their progress with REAP, which is a major step in going from a development system to an operational classroom system.

We sought a more quantitative measurement of acceptance, and so we examined the e-mail correspondence from the curriculum supervisor for the classes in which REAP is used. In the previous semester, starting in September 2005, there were frequent complaints about document quality and errors in the system. For example, one such e-mail noted that in some of the documents, "the [vocabulary] items are not complete sentences." We therefore decided to examine the frequency of certain words in her E-mails over time. We defined a set of "BAD" words that occurred often in complaints about the system (e.g., "glitch," "problem," "terrible," "inappropriate," "bad," "worry," "terrible"), as well as a smaller set of "GOOD" words that indicate acceptance of the system (e.g., "interesting," "appreciate," "better," "nice," "good," "helpful," "thanks," "ok"). We also examined the occurrence of any negative, or "NOT", words (e.g., "not ," "never," "can't") that appeared often in complaints about the system (e.g., "…didn't work," "Document 28047 didn't show up at all"). The frequency of both negative words and BAD words decreased over time from last semester until the present time, which corresponds to the decreasing number of complaints about the system by the ELI. The frequency of GOOD words increased over the same period. We also checked the frequency of the word "the" over time to validate these results, and it stayed fairly constant at around five percent of words, as we expected. The selection of these words was somewhat arbitrary and ad hoc, but we did not avoid or

remove any words. Neither did we exclusively examine earlier e-mails to find bad words, or later e-mails to find good words. Therefore, while these results certainly do not provide conclusive proof of acceptance, we feel that they strongly indicate that our system has improved significantly. The curriculum supervisor, who was not informed that we would use the E-mails for this purpose until after the analysis was complete, also agrees verbally that the system has improved a great deal. A graph of frequency over time for the three defined word types is shown in Figure 3. The horizontal axis shows the month and day in 2005-2006. The vertical axis corresponds to the type frequency, defined by the number of words in an E-mail that fall into each category divided by the total number of words in that E-mail. Polynomial lines of best fit show the trends in the data.

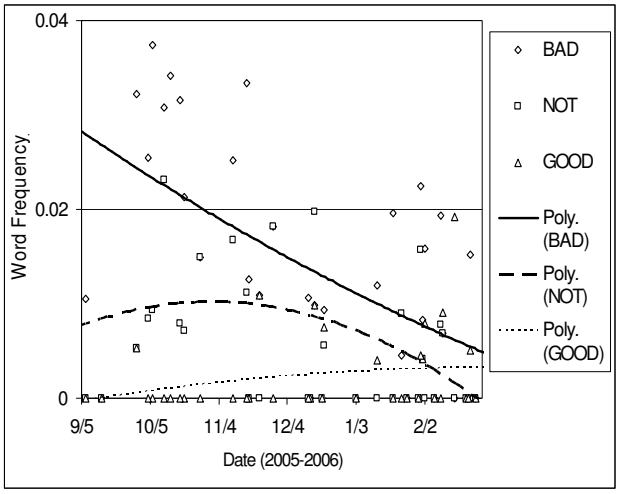

Figure 3 *Graphs showing the trends of frequency over time of "GOOD," "BAD," and negative words*

The opinions of the students at the ELI also affected its acceptance into the classroom. In addition to the exit survey discussed previously, the students are asked after each reading to fill out an online survey about the difficulty and interest of the passage they just read. They respond on a Likert scale, with a value of one corresponding to the least difficulty or interest, and a value of five corresponding to the most difficulty or interest. While the difficulty feedback ratings are near the ideal middle value of three (3.10), the interest ratings are also near three (3.08). Students appear to find the reading passages appropriately difficult, but not always engaging. Figure 4 shows graphs of the ratings.

4. Conclusion

The REAP system has satisfied a number of criteria in order to gain acceptance into the classroom at the English Language Institute at the University of Pittsburgh. REAP presents useful web documents of the right difficulty level in a way that is conducive to learning and can be integrated into the ELI curriculum. Some issues are yet to be resolved, such as the topic and context of readings. These issues will be addressed in future versions of the system as it moves from focusing solely on teaching single words toward teaching word collocations and grammar.

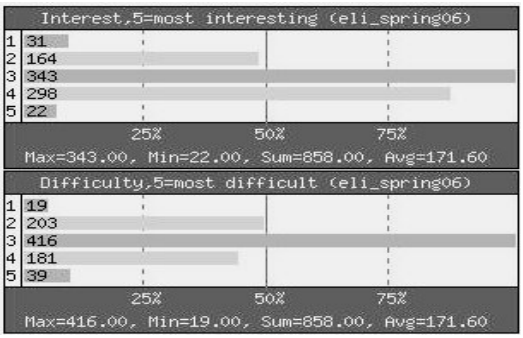

Figure 4: *Post-Reading Difficulty and Interest Feedback Ratings by Students*

5. Acknowledgments

The authors thank Jon Brown and James Sanders for their work on the REAP project. We also thank Alan Juffs and Lois Wilson at English Language Institute at the University of Pittsburgh for using REAP in the classroom.

This material is based on work supported by NSF grant IIS-0096139. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

6. References

[1] Brown, J. and Eskenazi, M. (2004) "Retrieval of authentic documents for reader-specific lexical practice." In Proceedings of InSTIL/ICALL Symposium 2004. Venice, Italy.

[2] Collins-Thompson, K. and Callan, J. (2004) "Information retrieval for language tutoring: An overview of the REAP project" (poster description). In Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK.

[3] Ogilvie P. and Callan, J. (2002) Experiments using the Lemur toolkit, Proceedings of the 2001 Text REtrieval Conference, NIST special publication 500-250: 103-108.

[4] Collins-Thompson, K. and Callan, J. (2004) "A language modeling approach to predicting reading difficulty." In Proceedings of the HLT/NAACL 2004 Conference. Boston.

[5] Klein, D. and Manning, C. D. (2002) Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), December 2002.

[6] Coxhead, A. (2000) A New Academic Word List. TESOL Quarterly, 34(2): 213-238.

[7] Walter, E., editor. (2005) *Cambridge Advanced Learner's Dictionary, 2nd Edition*. Cambridge University Press.

[8] Brown, J., Frishkoff, G., and Eskenazi, M.. (2005). "Automatic question generation for vocabulary assessment." In Proceedings of HLT/EMNLP 2005. Vancouver, B.C.