

Language Learning: Challenges for Intelligent Tutoring Systems

Michael Heilman

Language Technologies Institute,
Carnegie Mellon University
mheilman@cs.cmu.edu

Maxine Eskenazi

Language Technologies Institute,
Carnegie Mellon University
max@cs.cmu.edu

Abstract. We describe the challenges presented by the assessment and presentation of knowledge components in the language learning domain, with particular attention to vocabulary acquisition. This paper first discusses the fact that the meaning of words is not as well formalized as many topics in better-defined domains such as mathematics. There follows a comparison of the number of knowledge components in language to the much lower amounts in other domains. An Intelligent Tutoring System for language must adopt different presentation and assessment strategies to confront the specific challenges of the domain. We describe REAP, a system that confronts these issues, and present empirical results that demonstrate its effectiveness.

Keywords: Intelligent Tutoring Systems, Computer-Assisted Language Learning

INTRODUCTION

Language learning is a multi-level task that integrates elements such as words, syntax, pronunciation, and culture. For one part of language learning, learning how to read, it is very different from more structured domains in that there are tens of thousands of knowledge components to be learned rather than a few hundred. In mathematics, knowledge components consist of particular formulae and theorems along with methods for their application. In learning to read, however, the set of knowledge components includes not only all of the grammatical rules in a language as well as exceptions to those rules, but also all of the lexical items in a language. Dictionaries, vocabulary lists, and other lexical resources define word meaning in an informal and limited capacity, and are not well suited for direct study. Knowledge of each word has not, or perhaps cannot, be defined as explicitly and formally as in other domains. Different teaching and assessment strategies must therefore be employed.

Description of the REAP Tutoring System

We begin with a presentation of the REAP reading tutoring system (Collins-Thompson and J. Callan, 2004 and Brown and Eskenazi, 2004), the development of which fuels our discussion of the many challenges of language tutoring. The goal of the REAP system is to furnish appropriate, authentic texts to students to help in reading and vocabulary learning. The tutoring system will incorporate grammatical constructions in the near future, but until now has focused primarily on teaching vocabulary. In REAP, a student sees short reading passages that contain a number of words (usually ranging from two to four) from his or her list of target words to be learned from context. The passages are Web documents of about one to two pages in length, covering a wide variety of topics.

In the current iteration of the system, a student user has a list of target words that he or she needs to learn over the course of a semester. We need to generate a list of words that appear in documents of the proper grade level but which the student has not previously learned. The student takes a preliminary test in which one question is presented for each word in a list of words that are assumed to be just above reading level for that student. This method takes a great deal of time, and brings up several issues related to assessment that we will discuss below.

Finding and identifying appropriate documents for these reading passages is also a significant challenge because of constraints on length, readability, topic, context, and text quality, as well as the preference for documents that contain multiple target words. We have found that less than one percent of the documents containing any target words are actually suitable for the students when we use the above-mentioned constraints. We will discuss the criteria for finding good documents in greater detail below.

The target words in a reading passage are highlighted to draw attention to them. In addition, students can look up these target words, as well as any other unknown words in the passage, by using an electronic version of the Cambridge Advanced Learner's Dictionary (Woodford and Jackson 2004) integrated into REAP. All dictionary use by students is tracked.

After each reading passage, the student works through exercises that facilitate construction and refinement of knowledge components for the target words. In later sections we discuss the various problems we have encountered that are related to the creation and evaluation of these exercises. These exercises are also a way to assess student knowledge, and can be used to select the subsequent reading material. It is difficult to accurately assess vocabulary knowledge, however, because of various issues that are specific to the language domain. We discuss these assessment-related challenges in the next section.

ISSUES RELATED TO STUDENT KNOWLEDGE OF VOCABULARY

A major issue we have encountered while creating this tutoring system centers on the assessment of a student's knowledge of words. This is essential in order to present readings that sufficiently challenge the student and provide effective instruction. Before choosing documents to present, the system must have an idea of what the student needs to learn. The REAP system therefore presents a pre-test to determine which words, from a chosen list, the student does not know. Once the tutoring begins, if a student has already learned a target word from a prior reading, then it is not efficient to next present a document with that word. Conversely, if a student has not learned a word from several prior readings, then it may not be worthwhile to present a document with that word in a new context. Also, it is important to have a model of the student's overall vocabulary knowledge. If a student's overall vocabulary knowledge is overestimated, the system will consistently search for readings that are too difficult and impede learning. Conversely, if the vocabulary is underestimated, the system will search for readings that are not sufficiently challenging.

Considering individual words, beyond the morphological relations between some words e.g., “select”, “selection”, “selecting”), there is little or no overlap among word usage patterns that allows for prediction of knowledge from synonyms or related words. Two semantically related words may appear in very different contexts. For instance, it is difficult to accurately predict whether a student knows the meaning of “industrious” from the fact that he or she knows the meaning of “hard-working”. Also, language courses cover hundreds of words—and beyond the classroom students learn thousands more; testing all these words is not usually feasible. In contrast, a course in a well-defined domain such as mathematics may have a curriculum consisting of fifty knowledge components, making it easier to test each and every one of them reliably.

Another reason for the difficulty of assessing individual words is that there are multiple levels of knowing a word. It is easier for a student to recognize the meaning of a word in a sentence than to produce a sentence of his or her own using that word. Stahl (1986) proposed a model of word knowledge with three levels with various degrees of knowledge. In the first level, a student is unfamiliar with the word. In the second level, the student has passive knowledge of the word and can understand it when reading or listening but cannot produce it. In the third level, the student has active knowledge and can produce the word in novel contexts. We assess passive word knowledge in REAP with a variety of multiple choice questions (Brown, et al. 2004). Synonym and antonym questions are generated by using WordNet (Fellbaum 1998) in conjunction with frequency statistics so that neither overly rare nor common words appear in exercises. Cloze questions, examples of which are shown in Figure 1, are generated automatically as well by extracting passages that contain the target word in an informative context. Finally, students are asked to produce novel sentences demonstrating knowledge of words. Thus REAP generates exercises that assess the various levels of knowledge of a word, from passive to active. The sentence production items currently have to be hand-graded, and are used only as post-test items.

He could never ___ the success he had enjoyed with his first record.
acknowledge comprise induce reproduce

Recently, the software company became a(n) ___ of a large corporation.
index subsidiary transmission interval

He answered the first question correctly, though he got all ___ questions wrong.
subsequent empirical identical legal

Figure 1: Example multiple-choice cloze questions generated automatically in REAP

A student may also know a word's meaning but not the set of words with which it is used conventionally. Words often occur in set phrases and also in collocations, which are pairs of words that co-occur more frequently than would be expected from semantic constraints alone. With collocations, the meaning of words is not necessarily compositional, such that the collocation "white wine" does not refer to wine that is white but rather yellow-colored wine made in a certain way. Students thus may know the individual meanings of words, but then use them improperly. For example, a non-native speaker might describe tea as "powerful" and a car as "strong,"

whereas a native speaker would assign the adjectives in the opposite way though they are basically synonymous. (Halliday, 1966). These collocations are difficult to identify in a tutoring system because there is no comprehensive list available electronically. However, collocations can be identified automatically by employing statistical measures of co-occurrence (Church and Hanks 1989), including the Chi-square statistic, likelihood ratio, and mutual information of two words. These measures can be calculated for a given pair of words from a corpus of text using the frequencies of co-occurrence of these words within a small window (a few words long), the frequencies of the words separately, and the total number of words in the corpus. For instance, using the REAP corpus of Web documents, the likelihood ratio of “exceedingly difficult” is 61.9, while the ratio for “usually difficult” is a non-significant 4.1. This indicates that the former is a collocation while the latter is not. We have found that useful collocations usually have values over 50 for the likelihood ratio statistic. We have developed a prototype version of REAP which highlights significant collocations that have been identified by co-occurrence statistics and part of speech information. The system also creates multiple choice questions to assess student knowledge of which pairs of words collocate and which occur together more or less by chance.

REAP as a Tool for Teachers and Researchers

It is often assumed that students at a given level have encountered and learned a set of words associated with that level. The teacher assumes that an intermediate student knows words like "say" and "give" without explicitly testing these words. Of course, presumption of prior knowledge occurs in any domain--calculus students should know long division, for example--but it is usually not possible for students in such domains to reach the given level without that prior knowledge. It is very likely, however, that an advanced student of a foreign language might have "gaps" or "holes" in his vocabulary. For instance, a word like “dinosaur,” which is known by any first grade student, might be unknown to a second language learner because it was never encountered in any lesson. Electronic dictionary use by students using the REAP tutoring system provides evidence of these "gaps" in student vocabulary knowledge. Although these students are at a language learning level approximately equivalent to eighth grade in an American school, they often look up words that are commonly learned before sixth or even fourth grade. We used the Living Word Vocabulary (LWV) to define first language grade levels for words (Dale and O'Rourke, 1981). In the LWV, the grade levels assigned to a word is the grade in American schools by which most students know that word. A chart of the proportion of the total number of dictionary accesses for words of each LWV level is shown in Figure 2. The data do not sum to one because there is a small percentage of looked up words for which there is no level defined in the LWV. While the probability of a student looking up words increases with that word's grade level, lower level words occur much more frequently than the rarer high-level words. Most words in a document are therefore lower level words normally acquired by sixth grade by native speakers. As a result, more than a third of words looked up by students were fourth or sixth grade words according to the LWV list. This indicates that there are often gaps in the lexical knowledge of students. Thus, any estimation of vocabulary knowledge based on a subset of words will be prone to error.

In the REAP system, we have developed tools that allow teachers and researchers to track the gaps in student vocabulary. We integrated Automatically updated reports show teachers which words that students are looking up while reading, and in which documents these words are looked up. For a given student or for a whole class, teachers and researchers can easily compare data on looked-up words to performance on post-reading exercises, the time spent per document by a student, or any other data that are tracked in the REAP system. These tools allow teachers to track the gaps in student vocabulary and address them in class. Researchers can also look for patterns in the gaps that might correlate with the native language of the students or other factors.

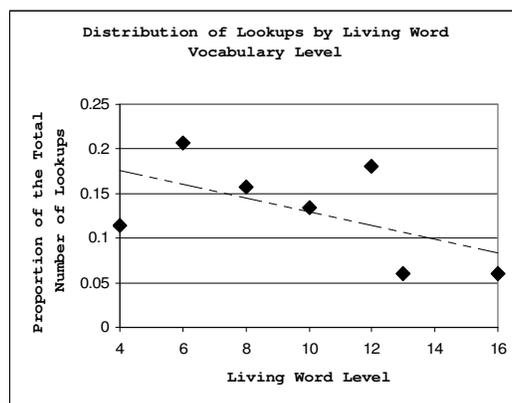


Figure 2: Proportion of total lookups by Living Word Vocabulary level

We also employ various heuristics in assessing student knowledge of individual target words. It is unclear how often, at what times, or in how many different contexts a word must be presented in order for a student to learn that word properly, though some research has been done on the topic (Pavlik, 2005 and Zahar, 2001). To refine word knowledge, we feel it is important to present words multiple times in authentic documents with various contexts. It has been shown that students acquire better knowledge of words when they are presented in multiple contexts (Zahar, et al., 2001). We chose to present each target word three times in order to facilitate knowledge refinement. There are many alternatives to such a scheme. In addition, it is unclear what exercise types most accurately assess student knowledge of words. The REAP system provides a way for researchers to determine the optimal number and timing for presentation of new vocabulary, as well as the best assessment strategies for vocabulary.

CHALLENGES OF PRESENTATION

The proper presentation of words is a major issue we have encountered in the REAP system. There are thousands of words that a language student must learn. Short of asking them to memorize lists of words and definitions, it is not feasible to present each individually. A language tutoring system must attempt to teach multiple items in the same reading passage, taking into account the uncertainties about what a student knows that result from the assessment problems detailed above. We will discuss this issue from the point of view of vocabulary teaching, although it applies to grammar as well.

When multiple words must be presented in the same reading, it is important to know the optimal number of unknown words to present in any given passage—which is called the vocabulary "stretch". If too many words in a reading are unknown, then the student is likely to become confused and learn none of them since it would be hard to understand the meaning of the passage. Research has indicated that a reader must know 98% of the words in a passage for it to be comprehensible without access to a dictionary (Hsueh-chao, et al. 2000). It is unclear what percentage of words must be known if a dictionary is available, and whether this threshold depends on the topic or individual. Also, it is not clear what percentage of words must be known in order for a student to learn vocabulary from contextual clues alone. A current study using REAP is aimed at clarifying some of these issues.

In our tutoring system, we use a readability measure based on language modeling techniques (described in Collins-Thompson and Callan 2004) in order to find documents at the appropriate level of difficulty. Collins-Thompson and Callan gathered a corpus of documents labeled with reading level for first through twelfth grade levels (in an American school). The unigram frequencies of individual words, as well as the rate of out of vocabulary words at each grade was used to identify the words that appear frequently and infrequently in texts at certain grades. The frequencies of words in a new text can then be compared to the models for each of grades in order to identify which grade level it is most likely the new text corresponds to. The accuracy of this measure compares favorably to that of other readability measures, especially on Web documents. Even though the REAP system can fairly accurately predict the reading level of a document, the number of words that students look up in a document varied from zero to thirty or more per reading in a current study. In the study, the mean number of words looked up was about 4.0, and the standard deviation was 3.5. Finding documents that contain the optimal number of unknown items while adhering to other criteria, such as reading level, is a challenge.

We can estimate the percentage of unknown words in a document by summing the number of target words and other words that a student looked up, and dividing by the total number of distinct word tokens. These estimates show that the REAP system is fairly successful at selecting passages of appropriate difficulty because for most documents less than three percent of the words are either target words or additional unknown words the student looked up while reading the document. In addition, student feedback gathered after each reading indicates that students find the documents neither too difficult nor too easy.

The optimal passage length for a reading is another issue for REAP. Currently, we present documents of approximately a thousand words to students. These documents provide a great deal of context in the hopes that students will acquire robust knowledge of the new vocabulary items. It may, however, be more effective to present new words with less contextual information. A single paragraph, or a single sentence, may be sufficient for a student to learn a word accurately. Presenting words in shorter passages would probably allow for greater control and efficiency in moving the student through a curriculum, but these shorter passages may not provide enough surrounding context for the words to be learned accurately. The REAP system will allow us to test these hypotheses in the near future.

Drawing focus to target vocabulary words affects student learning of those words. When presenting a word in a passage of any length, the student may not focus on and learn the target word if he or she is not prompted to do so. A student may very well skip over a target word because it is unknown and not necessary for comprehension of the passage. Highlighting target words is a way of increasing the student's 'noticing' of those words, which

has been shown to be very important in second language learning (Schmidt, 1990). Research has shown mostly positive effects of highlighting target words (De Ridder, 2002), and REAP has highlighted target words. This choice to draw attention to target words may have negative effects, of course. For instance, the student may choose to focus solely on the highlighted words and ignore the rest of the reading. In practice, we find that this is sometimes the case. Some students go through long documents in a few minutes because they only look at the highlighted target words. There are two problems that arise from students focusing only on highlighted words in this way. First, the surrounding context, which may be crucial to learning nuances of word meaning, is ignored. Second, there are likely a number of other new words in a passage that the student has the opportunity to learn. Although it is desirable to guide a student to focus on certain parts of a reading passage and on specific items, learning opportunities may be missed if the rest of the passage is ignored.

Dictionary accessibility is related to the prior issue of drawing attention to target words. In the REAP system, students can click on highlighted words to see a definition, and can also easily look up any other unknown words from the reading. Dictionary access can also lead to a student using only the dictionary definition to learn a word's meaning, instead of looking at the surrounding context as well. We feel, however, that it is more valuable for a student to be able to engage in a form of exploratory learning by looking up meanings for non-target words. Also, the dictionary allows the student to better grasp the surrounding context of a target word if that context contains unknown words as well. Students ignoring contextual information in favor of dictionary definitions can be seen as a form of "gaming the system," which is a common problem for intelligent tutoring systems (Baker, et al., 2004). Dictionary use and the highlighting of target words can have negative effects, but student behavior can be monitored either by teachers or automatically, to stop students from "gaming the system" and missing opportunities for learning.

Quality of Authentic Passages as Reading Material

Independent of passage length and difficulty, not all documents are of equal pedagogical value. The REAP system focuses on using authentic materials because they both improve student motivation and help in overcoming the cultural barriers to language acquisition (Bacon and Finnemann, 1990). In order to find a sufficient number of passages automatically, the system uses the Web as a corpus from which authentic reading materials can be gathered, but the quality of Web documents as readings is a crucially important issue. Many documents on the Web consist of long lists of names or words, with few well-formed sentences. Common sense tells us that such documents without well-formed sentences are not valuable as readings. Documents consisting primarily of coherent and cohesive sentences and paragraphs are generally much more engaging and valuable. Human filtering and selection of reading material is possible for a small-scale tutoring system, but in REAP all filtering is automatic. The first person to see much of the material used in REAP is the student. Automatic filtering of authentic materials greatly decreases the ratio of development time to instructional time, which is an important issue for intelligent tutoring systems.

We have implemented a text quality predictor in REAP that effectively filters out the large number of documents that are useless as reading material. At first, we attempted to identify useful documents by analyzing HTML structure in order to tell how much text might be contained in tables and lists of items. Examining the widely varying structure of Web pages, however, is not feasible because there is no set of consistent formatting criteria for Web pages. Instead, we base our measure of text quality on the confidence scores from a probabilistic natural language parser that produces parse trees representing grammatical structure for each sentence in a document. These sentence-level confidence scores are estimates of the log likelihood of the most likely parse tree for the given sentence. Lists, menus, and other features that appear often in poor quality documents consist of series of noun phrases for which there is no likely syntactic parse tree (e.g., "electronics cameras computers games appliances mobile phones"). Likewise, other poor reading material such as bulletin board postings often contain incomplete sentences for which there is no likely parse. We use the Stanford parser to generate these sentence-level parser confidence scores (Klein and Manning, 2002). Significantly fewer than half of the documents gathered during crawls of the Web pass the text quality filter, illustrating that good reading passages for students are difficult to find and identify automatically. Of course, the cohesiveness and quality of contextual information is something that is difficult to define quantitatively, and so we had to set a threshold for the level of quality that is acceptable for a document to be presented to students.

Student interest in reading material is another important issue in language learning. Prior research has shown that personalization and choice in tutoring systems can facilitate learning in other domains (Cordova and Lepper, 1996). While tutoring systems for well-defined domains such as mathematics have little control over the context in which material is presented to students, language learning material can be taught in almost any context. The great majority of words (grammar as well) is not specific to any given topic or context. So while an algebra tutoring system might be forced to present a given theorem in the context of a business transaction, a language tutoring system can present a word such as "specific" in any context, for example in passages referring to "a specific sports team," "a specific food," "a specific car," etc. The great variety of topics in which language

teaching may be situated is a great advantage to tutoring systems. While a human teacher usually gives a single reading that may interest only a subset of students, a computer tutoring system can provide individual instruction tailored to each student's interests. Although our system does not currently incorporate topic detection and tracking, we feel that personalization and choice are important goals for any language tutoring system. We have created a text categorization system based on Support Vector Machines (Burgess 1998) using SVM-Light (Joachims 2002) that assigns one of ten topic labels from the top level of the topic hierarchy of the Open Directory Project (ODP, <http://dmoz.org>) with 78% accuracy. The topic labels include "Arts," "Science," "Business," and others. The system was trained and tested on a set of 10,000 labeled documents gathered from the ODP in early 2006. The corpus was split randomly into a training set of 8,000 documents and a test set of 2,000 documents. In an upcoming study, REAP will use these topic labels to provide documents to students according to their individual interests.

Automatically finding appropriate reading material is not a trivial task as one might at first assume. Although finding arbitrary documents that contain single target words is simple with modern search engine technology, these documents are rarely useful as reading passages. It is even more difficult to find pairs or groups of specific words since the likelihood of rare words occurring together is so low. For example, if two words each occur in one in a thousand documents, then unless they are strongly related they will occur together in about one in a million documents. What further complicates the matter is that most documents do not satisfy length, readability, and text quality constraints either—not to mention topic constraints. In generating our database of reading passages, we found that only about 0.5% of documents pass through our filters. More than half of documents are too long, about a third are out of the appropriate range of estimated reading difficulty, and about two-thirds of documents consist mainly of lists and menus rather than cohesive text. It is therefore very difficult for an intelligent tutoring system to select appropriate reading passages to present to students, even independent of the nature of the presentation.

EMPIRICAL RESULTS FROM A STUDY INVOLVING REAP

In this section we will present results from a recent study that validate the approach used in the REAP system. The study was conducted in the Spring of 2006 at the English Language Institute of the University of Pittsburgh. Thirty-two English as a Second Language (ESL) students used the system once a week in eleven forty-minute sessions over the course of the semester. Twenty-two of these students successfully completed the post-test. Some did not show up for the post-test because it was voluntary rather than for a grade, and a few experienced minor technical difficulties in one portion of the test, so their scores were excluded from the results we present. The subjects were international students from a variety of countries—including Saudi Arabia, Korea, Japan, and France—who were studying English in order to enter American universities. The students were at an intermediate ESL level corresponding to about eighth grade in a U.S. elementary school.

The REAP system supplemented their coursework in an English reading skills course. A list of 216 target vocabulary words was created for the study. These words were chosen from the Academic Word List (Coxhead 2000), which consists of words that are essential for reading and writing University-level English text. These words are normally learned at a level above the ESL level of the students, and none appeared in the course's regular materials. It is thus unlikely, although still possible, that the students would learn these words during the semester outside of the REAP system. Each student took a 45 minute pre-test consisting of multiple choice cloze (that is, fill-in-the-blank) exercises in order to assess which of these words they did and did not know. Most students completed about half of the list of possible pre-test items. The words for which responses were incorrect were added to individual student focus word lists, on which the REAP system would provide training.

During sessions with REAP, the students read Web documents selected by REAP containing up to four of the words from a particular student's focus word list. These documents also passed the various automatic filters in REAP for text quality, length, and reading level. The REAP system attempted to show each focus word up to three times in three different documents. As discussed above, students were able to look up any additional unknown words using an electronic dictionary. After each reading, students completed multiple-choice cloze exercises related to focus words from the document just read. In most but not all cases, students received training on their entire list of focus words.

At the end of the semester, one week following the final reading session, students took a post-test to assess their progress made while using REAP. The post-test had two sections. In the first section, students were asked to produce novel sentences demonstrating knowledge of the meaning of a word for ten of the focus words on which they received training. The second part of the post-test consisted of forty previously unseen multiple-choice cloze exercises that were similar to the pre-test and post-reading exercises. The first part of the test—the goal of which was to assess transfer of knowledge to a novel task—was conducted before the other section so that the cloze exercises did not give away sentences which could be used to answer the transfer items. In a separate test conducted a few days later, eight of the students were asked to produce sentences demonstrating

knowledge of looked-up words. The words for this test were not on a student’s focus word list, but instead were words looked up in the electronic dictionary while the student used REAP. All of the sentence production items were graded using the following three-point system: one point was given for proper grammar usage of the word, one point was given for the word’s semantic meaning fitting into the sentence regardless of whether knowledge was demonstrated (e.g., “The student *demonstrated* his success.”), and a third point for clearly demonstrating knowledge of the word (e.g., “The student *demonstrated* his knowledge to the teacher by writing a sentence”).

The results from the post-test for focus words are presented in Figure 3. The pre-test scores for all of these words are essentially zero because students answered pre-test exercises for these words incorrectly. For the pre-test multiple-choice cloze exercises, there was a twenty-five percent chance of guessing the answer correctly since there were four possible choices. For the sentence production tasks, the chance of producing a valid answer was considerably lower, although if the part of speech of the word was known, a student might receive partial credit for a grammatical but meaningless sentence. The students performed very well on the multiple-choice cloze items, indicating that using REAP helped them to learn their focus words. Although performance on the transfer task was lower, there is some evidence that the knowledge gained on focus words while using REAP transferred to a novel situation. Teachers of the course predicted low performance on the sentence-production transfer task beforehand due to the level of the students and the comparative difficulty of production tasks over recognition tasks.

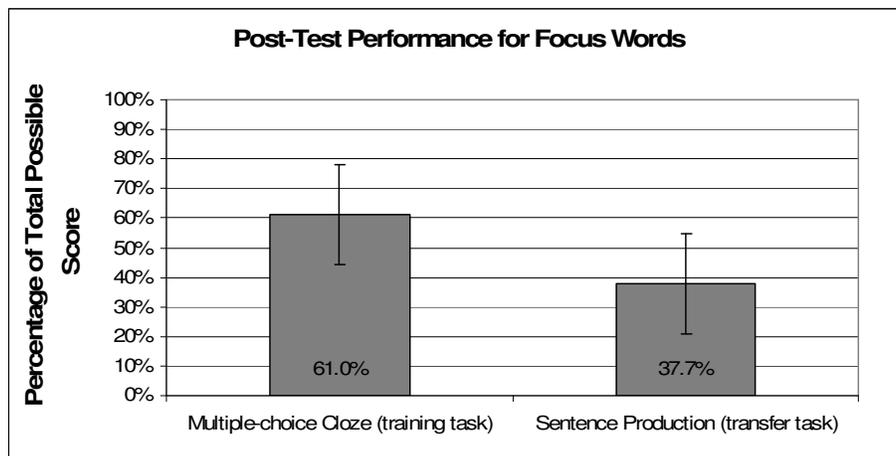


Figure 3: Student post-test performance on focus words for the training task and a transfer task

The production scores for looked-up words are presented for comparison with the scores for focus words in Figure 4. From the scores for looked-up words, students appear to have learned a small but significant portion of the words beyond their individual focus word lists. Several students looked up more than one hundred additional words during reading, so even if a small percentage of these words were learned, it would be valuable for accelerated learning.

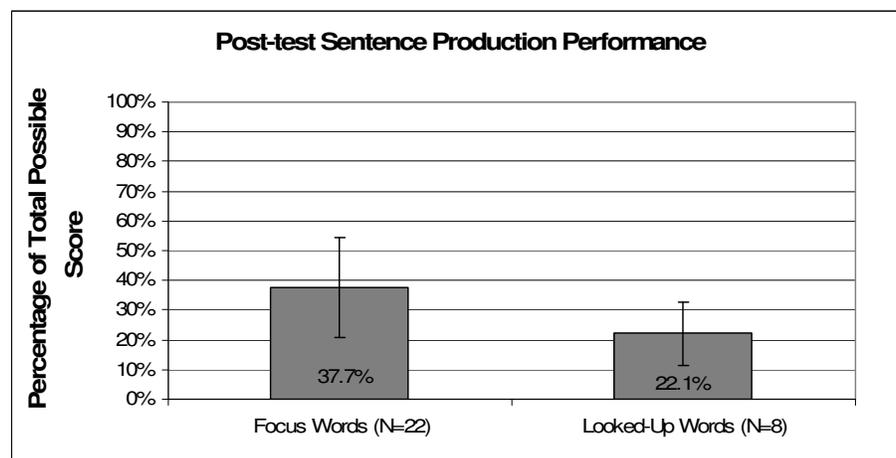


Figure 4: Sentence production performance for focus words compared to looked-up words

The students' opinions of the system also validate REAP. Prior to one of the final reading sessions, students took an exit survey asking ten questions on a Likert scale from one to five. A value of one indicated strong disagreement with a given statement, while a value of five indicated strong agreement. The results of this exit survey are shown in Figure 5, with bars for standard error. On the positive side, students found the system very easy to use, and most felt that REAP should be used in future classes. They also believed that they had learned a lot of the focus words, corroborating the quantitative evidence from the post-test. Also, the students did not feel that the documents selected by REAP were too difficult for them to understand, which indicates that our automatic filters work properly to provide high-quality, authentic documents. On the negative side of things, students did not find the documents particularly interesting, and would have liked to select the documents or at least the general topics of the documents. Overall, however, the exit survey results were very positive, as were the additional free-response comments of the students.

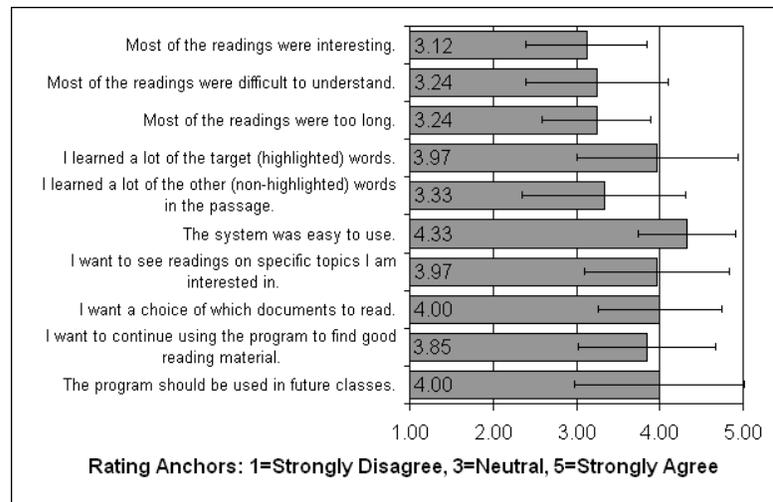


Figure 5: Exit Survey Results from 32 students who used the REAP system

CONCLUDING REMARKS

Language tutoring systems face a unique set of problems that arise from the nature of the language domain. The set of items, either grammatical or lexical, that a student must learn is very large. Also, it is often difficult to accurately assess the knowledge of a single item because of the various contexts in which words may occur. Choosing the number, nature, and timing of assessment exercises is thus a great challenge. In addition, knowledge components must often be presented together in a single reading passage because teaching items individually is not feasible. A language tutor must choose the optimal number of items to present in a passage, the length of the passage, and whether to draw attention to these items, in order to make learning efficient. When using authentic material, which is desirable in language learning situations to increase motivation and transfer to real-life situations, documents must also be filtered for text quality to identify useful passages that consist of cohesive sentences and paragraphs. Finally, language learning materials can cover a wide variety of topics such that there are unique opportunities for personalization and choice in a language tutoring system. Although similar issues often arise in other learning domains, their unique combination provides a number of interesting challenges for intelligent language tutoring systems.

The REAP system addresses many of these challenges of the language learning domain, and empirical results validate our approach. We are planning to develop better automatic question generation techniques, perhaps including techniques for evaluating free responses, so that we can better assess students' knowledge. It may also be useful for us to estimate the "holes" in students' vocabulary by using language background and dictionary access information. We are investigating the optimal length and scheduling of the passages that REAP shows to students. In the near future, the REAP will also include options for personalization and choice of reading topics in order to make students more interested and engaged while using the system. The system will eventually expand to include grammar instruction along with lexical practice.

ACKNOWLEDGMENTS

The authors thank Jamie Callan, Kevyn Collins-Thompson, 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. We thank Vincent Aleven for some early guidance for the paper, as well as the anonymous reviewers for providing useful feedback.

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.

REFERENCES

- Bacon, S., and Finnemann, M. (1990). A study of attitudes, motives, and strategies of university foreign language students and their disposition to authentic oral and written input. *Modern Language Journal*, 74, 459-473.
- Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". Proceedings of ACM CHI 2004: Computer-Human Interaction (2004) 383-390.
- Brown, J. and Eskenazi, M. (2004) "Retrieval of authentic documents for reader-specific lexical practice." In Proceedings of InSTIL/ICALL Symposium 2004. Venice, Italy.
- Brown, J., Frishkoff, G., and Eskenazi, M.. (2005). "Automatic question generation for vocabulary assessment." In Proceedings of HLT/EMNLP 2005. Vancouver, B.C.
- Burges C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2): 121-167.
- Church, K. W., and Hanks, P. (1989). Word Association norms, mutual information and lexicography. In *ACL* 27, pp. 76-83.
- 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.
- Collins-Thompson, K. and Callan, J. (2004) "A language modeling approach to predicting reading difficulty." In Proceedings of the HLT/NAACL 2004 Conference. Boston.
- Cordova, D. I., & Lepper, M. R. (1996) Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715-730.
- Coxhead, A (2000) A new academic word list. *TESOL Quarterly*, 34, 2: 213-238.
- Dale, E., and O'Rourke, J. (1976, 1981) *The living word vocabulary*. Chicago: World Book/Childcraft International.
- De Ridder, I. (2002) Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? *Language Learning and Technology* 6: 123-146.
- Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Halliday, M.A.K. (1994) Lexis as a linguistic level. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, and R.H. Robins (eds.), *In memory of J.R. Firth*, pp. 148-162. London: Longmans.
- Joachims, T. (2006). SVM light, An implementation of Support Vector Machines (SVMs) in C. <http://svmlight.joachims.org/>.
- 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.
- Open Directory Project (2006). <http://dmoz.org>.
- Pavlik, P. I. and Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559-586.
- Schmidt, R. (1990) The role of consciousness in second language learning. *Applied Linguistics*. 11, 129-158.
- Steven A. Stahl. (1986) Three principals of effective vocabulary instruction. *Journal of Reading*, 29.
- Woodford, K and Jackson, G. (2003). *Cambridge Advanced Learner's Dictionary*. Cambridge University Press.
- Zahar, R., Cobb, T., & Spada, N. (2001) Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Journal*, 57(4), 541-572.