Retrieval of Reading Materials for Vocabulary and Reading Practice

Michael Heilman, Le Zhao, Juan Pino and **Maxine Eskenazi**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{mheilman,lezhao,jmpino,max}@cs.cmu.edu

Abstract

Finding appropriate, authentic reading materials is a challenge for language instructors. The Web is a vast resource of texts, but most pages are not suitable for reading practice, and commercial search engines are not well suited to finding texts that satisfy pedagogical constraints such as reading level, length, text quality, and presence of target vocabulary. We present a system that uses various language technologies to facilitate the retrieval and presentation of authentic reading materials gathered from the Web. It is currently deployed in two English as a Second Language courses at the University of Pittsburgh.

1 Introduction

Reading practice is an important component of first and second language learning, especially with regards to vocabulary learning (Hafiz and Tudor, 1989). Appropriating suitable reading material for the needs of a particular curriculum or particular student, however, is a challenging process. Manually authoring or editing readings is time-consuming and raises issues of authenticity, which are particularly significant in second language learning (Peacock, 1997). On the other hand, the Web is a vast resource of authentic reading material, but commercial search engines which are designed for a wide variety of information needs may not effectively facilitate the retrieval of appropriate readings for language learners.

In order to demonstrate the problem of finding appropriate reading materials, here is a typical example of an information need from a teacher of an English as a Second Language (ESL) course focused on reading skills. This example was encountered during the development of the system. It should be noted that while we describe the system in the context of ESL, we claim that the approach is general enough to be applied to first language reading practice and to languages other than English. To fit within his existing curriculum, the ESL teacher wanted to find texts on the specific topic of "international travel." He sought texts that contained at least a few words from the list of target vocabulary that his student were learning that week. In addition, he needed the texts to be within a particular range of reading difficulty, fifth to eighth grade in an American school, and shorter than a thousand words.

Sending the query "international travel" to a popular search engine did not produce a useful list of results[1]. The first result was a travel warning from the Department of State[2], which was at a high reading level (grade 10 according to the approach described by (Heilman et al., 2008)) and not likely to be of interest to ESL students because of legal and technical details. Most of the subsequent results were for commercial web sites and travel agencies. A query for a subset of the target vocabulary words for the course also produced poor results. Since the search engine used strict boolean retrieval methods, the top results for the query "deduce deviate hierarchy implicit undertake" were all long lists of ESL vocabulary words[3].

We describe a search system, called REAP Search, that is tailored to the needs of language

[1] www.google.com, March 5, 2008
[2] http://travel.state.gov/travel/cis_pa_tw/cis_pa_tw_1168.html
[3] e.g., www.espindle.org/university_word_list_uwl.html

teachers and learners. The system facilitates the retrieval of texts satisfying particular pedagogical constraints such as reading level and text length, and allows the user to constrain results so that they contain at least some, but not necessarily all, of the words from a user-specified target vocabulary list. It also filters out inappropriate material as well as pages that do not contain significant amounts of text in well-formed sentences. The system provides support for learners including an interface for reading texts, easy access to dictionary definitions, and vocabulary exercises for practice and review.

The educational application employs multiple language technologies to achieve its various goals. Information retrieval and web search technologies provide the core components. Automated text classifiers organize potential readings by general topic area and reading difficulty. We are also developing an approach to measuring reading difficulty that uses a parser to extract grammatical structures. Part of Speech (POS) tagging is used to filter web pages to maintain text quality.

2 Path of a Reading

In the REAP Search system, reading materials take a path from the Web to students through various intermediate steps as depicted in Figure 1. First, a crawling program issues queries to large-scale commercial search engines to retrieve candidate documents. These documents are annotated, filtered, and stored in a digital library, or corpus. This digital library creation process is done offline. A customized search interface facilitates the retrieval of useful reading materials by teachers, who have particular curricular goals and constraints as part of their information needs. The teachers organize their selected readings through a curriculum manager. The reading interface for students accesses the curriculum manager's database and provides the texts along with support in the form of dictionary definitions and practice exercises.

3 Creating a Digital Library of Readings

The foundation of the system is a digital library of potential reading material. The customized search component does not search the Web directly, but rather accesses this filtered and annotated database of Web pages. The current library consists of approximately five million documents. Construction of the digital library begins with a set of target vocabulary words that might be covered by a course or set of courses (typically 100-1,500 words), and a set of constraints on text characteristics. The constraints can be divided into three sets: those that can be expressed in a search engine query (e.g., target words, number of target words per text, date, Web domain), those that can be applied using just information in the Web search result list (e.g., document size), and those that require local annotation and filtering (e.g., reading level, text quality, profanity).

The system obtains candidate documents by query-based crawling, as opposed to following chains of links. The query-based document crawling approach is designed to download documents for particular target words. Queries are submitted to a commercial Web search engine[4], result links are downloaded, and then the corresponding documents are downloaded. A commercial web search engine is used to avoid the cost of maintaining a massive, overly general web corpus.

Queries consist of combinations of multiple target words. The system generates 30 queries for each target word (30 is a manageable and sufficient number in practice). These are spread across 2-, 3-, and 4-word combinations with other target words. Queries to search engines can often specify a date range. We employ ranges to find more recent material, which students prefer. The tasks of submitting queries, downloading the result pages, and extracting document links are distributed among a dozen or so clients running on desktop machines, to run as background tasks. The clients periodically upload their results to a server, and request a new batch of queries.

Once the server has a list of candidate pages, it downloads them and applies various filters. The final yield of texts is typically approximately one percent of the originally downloaded results. Many web pages are too long, contain too little well-formed text, or are far above the appropriate reading level for language learners. After downloading documents, the system annotates them as described in the next section. It then stores the pages in a full-

[4] www.altavista.com

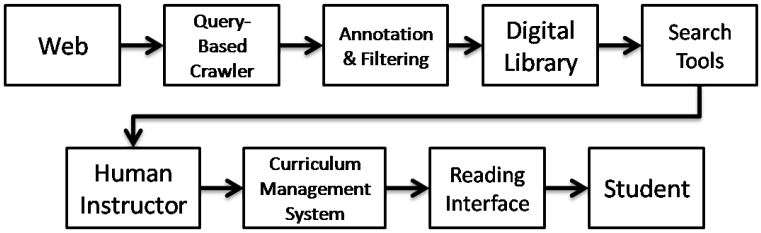

Figure 1: Path of Reading Materials from the Web to a Student.

text search engine called Indri, which is part of the Lemur Toolkit[5]. This index provides a consistent and efficient interface to the documents. Using Lemur and the Indri Query Language allows for the retrieval of annotated documents according to user-specified constraints.

4 Annotations and Filters

Annotators automatically tag the documents in the corpus to enable the filtering and retrieval of reading material that matches user-specified pedagogical constraints. Annotations include reading difficulty, general topic area, text quality, and text length. Text length is simply the number of word tokens appearing in the document.

4.1 Reading Level

The system employs a language modeling approach developed by Collins-Thompson and Callan (Collins-Thompson and Callan, 2005) that creates a model of the lexicon for each grade level and predicts reading level, or readability, of given documents according to those models. The readability predictor is a specialized Naive Bayes classifier with lexical unigram features. For web documents in particular, Collins-Thompson and Callan report that this language modeling-based prediction has a stronger correlation with human-assigned levels than other commonly used readability measures. This automatic readability measure allows the system to satisfy user-specified constraints on reading difficulty.

We are also experimenting with using syntactic features to predict reading difficulty. Heilman, Collins-Thompson, and Eskenazi (Heilman et al., 2008) describe an approach that combines predictions based on lexical and grammatical features. The grammatical features are frequencies of occurrence of grammatical constructions, which are computed from automatic parses of input texts. Using multiple measures of reading difficulty that focus on different aspects of language may allow users more freedom to find texts that match their needs. For example, a teacher may want to find grammatically simpler texts for use in a lesson focused on introducing difficult vocabulary.

4.2 General Topic Area

A set of binary topic classifiers automatically classifies each potential reading by its general topic, as described by Heilman, Juffs, and Eskenazi (2007). This component allows users to search for readings on their general interests without specifying a particular query (e.g., "international travel") that might unnecessarily constrain the results to a very narrow topic.

A Linear Support Vector Machine text classifier (Joachims, 1999) was trained on Web pages from the Open Directory Project (ODP)[6]. These pages effectively have human-assigned topic labels because they are organized into a multi-level hierarchy of topics. The following general topics were manually selected from categories in the ODP: Movies and Theater; Music; Visual Arts; Computers and Technology; Business; Math, Physics and Chemistry; Biology and Environment; Social Sciences; Health and Medicine; Fitness and Nutrition; Religion; Politics; Law and Crime; History; American Sports; and Outdoor Recreation.

Web pages from the ODP were used as gold-standard labels in the training data for the classifiers. SVM-Light (Joachims, 1999) was used as an implementation of the Support Vector Machines. In preliminary tests, the linear kernel produced slightly

[5] www.lemurproject.org

[6] dmoz.org

better performance than a radial basis function kernel. The values of the decision functions of the classifiers for each topic are used to annotate readings with their likely topics.

The binary classifiers for each topic category were evaluated according to the F1 measure, the harmonic mean of precision and recall, using leave-one-out cross-validation. Values for the $F1$ statistic range from .68 to .86, with a mean value of .76 across topics. For comparison, random guessing would be expected to correctly choose the gold-standard label only ten percent of the time. During an error analysis, we observed that many of the erroneous classifications were, in fact, plausible for a human to make as well. Many readings span multiple topics. For example, a document on a hospital merger might be classified as "Health and Medicine" when the correct label is "Business." In the evaluation, the gold standard included only the single topic specified by the ODP. The final system, however, assigns multiple topic labels when appropriate.

4.3 Text Quality

A major challenge of using Web documents for educational applications is that many web pages contain little or no text in well-formed sentences and paragraphs. We refer to this problem as "Text Quality." Many pages consist of lists of links, navigation menus, multimedia, tables of numerical data, etc. A special annotation tool filters out such pages so that they do not clutter up search results and make it difficult for users to find suitable reading materials.

The text quality filter estimates the proportion of the word tokens in a page that are contained in well-formed sentences. To do this it parses the Document Object Model structure of the web page, and organizes it into text units delineated by the markup tags in the document. Each new paragraph, table element, span, or divider markup tag corresponds to the beginning of a new text unit. The system then runs a POS tagger[7] over each text unit. We have found that a simple check for whether the text unit contains both a noun and a verb can effectively distinguish between content text units and those text units that are just part of links, menus, etc. The proportion of the total tokens that are part of content text units serves as a useful measure of text quality. We have found that a threshold of about 85% content text is appropriate, since most web pages contain at least some non-content text in links, menus, etc. This approach to content extraction is related to previous work on increasing the accessibility of web pages (Gupta et al., 2003).

[7]The OpenNLP toolkit's tagger was used (opennlp.sourceforge.net).

5 Constructing Queries

Users search for readings in the annotated corpus through a simple interface that appears similar to, but extends the functionality of, the interfaces for commercial web search engines. Figure 2 shows a screenshot of the interface. Users have the option to specify *ad hoc* queries in a text field. They can also use drop down menus to specify optional minimum and/or maximum reading levels and text lengths. Another optional drop-down menu allows users to constrain the general topic area of results. A separate screen allows users to specify a list of target vocabulary words, some but not all of which are required to appear in the search results. For ease of use, the target word list is stored for an entire session (i.e., until the web browser application is closed) rather than specified with each query. After the user submits a query, the system displays multiple results per screen with titles and snippets.

5.1 Ranked versus Boolean Retrieval

In a standard boolean retrieval model, with *AND* as the default operator, the results list consists of documents that contain all query terms. In conjunction with relevance ranking techniques, commercial search engines typically use this model, a great advantage of which is speed. Boolean retrieval can encounter problems when queries have many terms because every one of the terms must appear in a document for it to be selected. In such cases, few or no satisfactory results may be retrieved. This issue is relevant because a teacher might want to search for texts that contain some, but not necessarily all, of a list of target vocabulary words. For example, a teacher might have a list of ten words, and any text with five of those words would be useful to give as vocabulary and reading practice. In such cases, ranked retrieval models are more appropriate be-

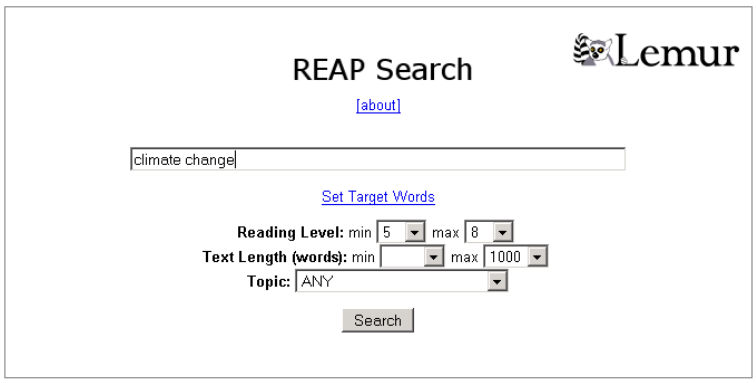

Figure 2: Screenshot of Search Interface for Finding Appropriate Readings.

cause they do not require that all of the query terms appear. Instead, these models prefer multiple occurrences of different word types as opposed to multiple occurrences of the same word tokens, allowing them to rank documents with more distinct query terms higher than those with distinct query terms. Documents that contain only some of the query terms are thus assigned nonzero weights, allowing the user to find useful texts that contain only some of the target vocabulary. The REAP search system uses the Indri Query Language's "combine" and "weight" operators to implement a ranked retrieval model for target vocabulary. For more information on text retrieval models, see (Manning et al., 2008).

5.2 Example Query

Figure 3 shows an example of a structured query produced by the system from a teacher's original query and constraints. This example was slightly altered from its original form for clarity of presentation. The first line with the *filrej* operator filters and rejects any documents that contain any of a long list of words considered to be profanity, which are omitted in the illustration for brevity and posterity. The *filreq* operator in line 2 requires that all of the constraints on reading level, text length and quality in lines 2-4 are met. The *weight* operator at the start of line 5 balances between the *ad hoc* query terms in line 5 and the user-specific target vocabulary terms in lines 6-8. The *uw10* operator on line 5 tells the system to prefer texts where the query terms appear together in an unordered window of size 10. Such proximity operators cause search engines to prefer documents in which query terms appear near each other. The implicit assumption is that the terms in queries such as "coal miners safety" are more likely to appear in the same sentence or paragraph in relevant documents than irrelevant ones, even if they do not appear consecutively. Importantly, query terms are separated from target words because there are usually a much greater number of target words, and thus combining the two sets would often result in the query terms being ignored. The higher weight assigned to the set of target words ensures they are not ignored.

6 Learner and Teacher Support

In addition to search facilities, the system provides extensive support for students to read and learn from texts as well as support for teachers to track students' progress. All interfaces are web-based for easy access and portability. Teachers use the search system to find readings, which are stored in a curriculum manager that allows them to organize their selected texts. The manager interface allows teachers to perform tasks such as specifying the order of presentation of their selected readings, choosing target words to be highlighted in the texts to focus learner attention, and specifying time limits for each text.

The list of available readings are shown to students when they log in during class time or for homework. Students select a text to read and move on to the reading interface, which is illustrated in Figure 4. The chosen web page is displayed in its original format except that the original hyperlinks and pop-ups are disabled. Target words that were

```
1  #filrej( #syn( PROFANITY HERE )
2     #filreq( #band(#greater(textquality 85)
3                    #greater(readinglevel 6) #less(readinglevel 9)
4                    #greater(doclength 300) #less(doclength 1000))
5                    #weight(1  #combine(business ethics) 1 #uw10(business ethics)
6                            10 #combine(motive amend manipulate mutual pursue
7                                  equivalent sole implement exploit neutral
8                                  utilize primary sector framework extract))))
```

Figure 3: Example Structured Query. The line numbers on the left are for reference only.

chosen by the teacher are highlighted and linked to definitions. Students may also click on any other unknown words to access definitions. The dictionary definitions are provided from the Cambridge Advanced Learner's Dictionary[8], which is authored specifically for ESL learners. All dictionary access is logged, and teachers can easily see which words students look up.

The system also provides vocabulary exercises after each reading for additional practice and review of target words. Currently, students complete cloze, or fill-in-the-blank, exercises for each target word in the readings. Other types of exercises are certainly possible. For extra review, students also complete exercises for target words from previous readings. Students receive immediate feedback on the practice and review exercises. Currently, sets of the exercises are manually authored for each target word and stored in a database, but we are exploring automated question generation techniques (Brown et al., 2005; Liu et al., 2005). At runtime, the system selects practice and review exercises from this repository.

7 Related Work

A number of recent projects have taken similar approaches to providing authentic texts for language learners. WERTi (Amaral et al., 2006) is an intelligent automatic workbook that uses texts from the Web to increase knowledge of English grammatical forms and functions. READ-X (Miltsakaki and Troutt, 2007) is a tool for finding texts at specified reading levels. SourceFinder (Sheehan et al., 2007) is an authoring tool for finding suitable texts for standardized test items on verbal reasoning and reading comprehension.

The REAP Tutor (Brown and Eskenazi, 2004; Heilman et al., 2006) for ESL vocabulary takes a slightly different approach. Rather than teachers choosing texts as in the REAP Search system, the REAP Tutor itself selects individualized practice readings from a digital library. The readings contain target vocabulary words that a given student needs to learn based on a student model. While the individualized REAP Tutor has the potential to better match the needs of each student since each student can work with different texts, a drawback of its approach is that instructors may have difficulty coordinating group discussion about readings and integrating the Tutor into their curriculum. In the REAP Search system, however, teachers can find texts that match the needs and interests of the class as a whole. While some degree of individualization is lost, the advantages of better coordinated support from teachers and classroom integration are gained.

8 Pilot Study

8.1 Description

Two teachers and over fifty students in two ESL courses at the University of Pittsburgh used the system as part of a pilot study in the Spring of 2008. The courses focus on developing the reading skills of high-intermediate ESL learners. The target vocabulary words covered in the courses come from the Academic Word List (Coxhead, 2000), a list of broad-coverage, general purpose English words that frequently appear in academic writing. Students used the system once per week in a fifty-minute class for eight weeks. For approximately half of a session, students read the teacher-selected readings and worked through individualized practice exercises.

[8] dictionary.cambridge.org

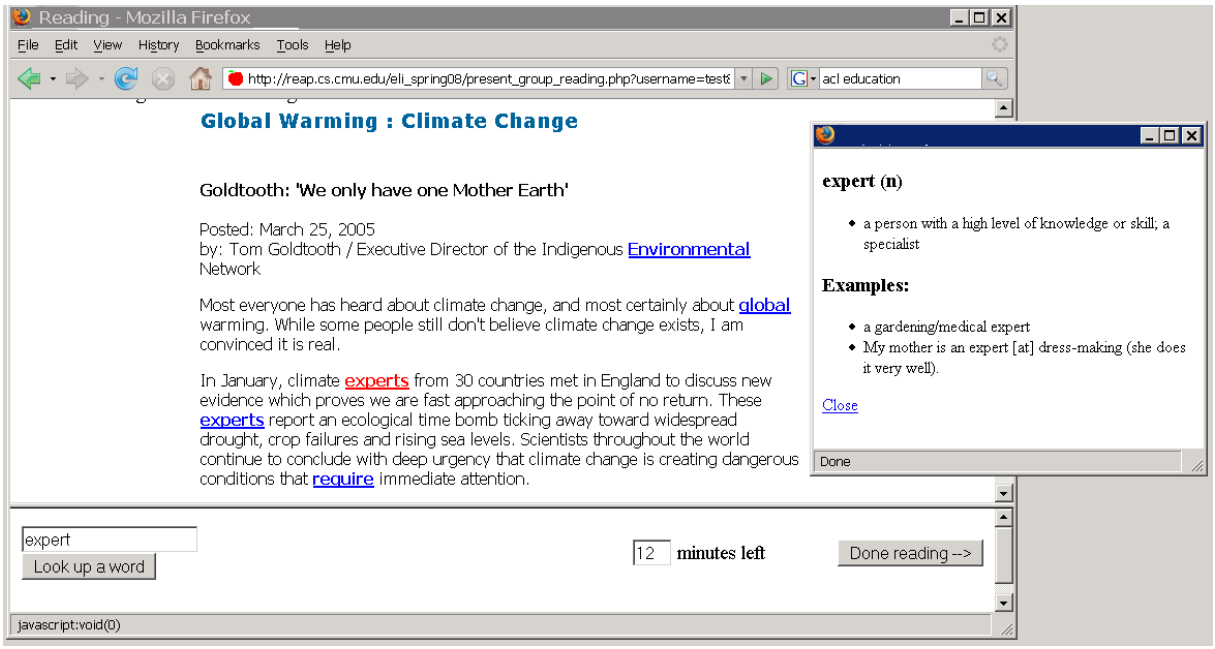

Figure 4: Screenshot of Student Interface Displaying a Reading and Dictionary Definition.

For the other half of each session, the teacher provided direct instruction on and facilitated discussion about the texts and target words, making connections to the rest of the curriculum when possible. For each session, the teachers found three to five readings. Students read through at least two of the readings, which were discussed in class. The extra readings allowed faster readers to progress at their own pace if they complete the first two. Teachers learned to use the system in a training session that lasted about 30 minutes.

8.2 Usage Analysis

To better understand the two teachers' interactions with the search system, we analyzed query log data from a four week period. In total, the teachers used the system to select 23 readings for their students. In the process, they issued 47 unique queries to the system. Thus, on average they issued 2.04 queries per chosen text. Ideally, a user would only have to issue a single query to find useful texts, but from the teachers' comments it appears that the system's usability is sufficiently good in general. Most of the time, they specified 20 target words, only some of which appeared in their selected readings. The teachers included *ad hoc* queries only some of the time. These were informational in nature and addressed a variety of topics. Example queries include the following: "surviving winter", "coal miners safety", "gender roles", and "unidentified flying objects". The teachers chose these topics because they matched up with topics discussed in other parts of their courses' curricula. In other cases, it was more important for them to search for texts with target vocabulary rather than those on specific topics, so they only specified target words and pedagogical constraints.

8.3 Post-test and Survey Results

At the end of the semester, students took an exit survey followed by a post-test consisting of cloze vocabulary questions for the target words they practiced with the system. In previous semesters, the REAP Tutor has been used in one of the two courses that were part of the pilot study. For comparison with those results, we focus our analysis on the subset of data for the 20 students in that course. The exit survey results, shown in 5, indicate that students felt it was easy-to-use and should be used in future classes. These survey results are actually very similar to previous results from a Spring 2006 study with the REAP Tutor (Heilman et al., 2006). However, responses to the prompt "My teacher helped me to learn by discussing the readings after I read

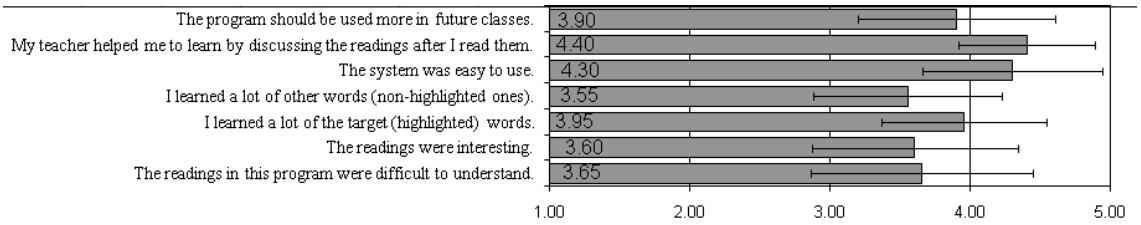

Figure 5: The results from the pilot study exit survey, which used a Likert response format from 1-5 with 1=Strongly Disagree, 3=Neither Agree nor Disagree, and 5=Strongly Agree. Error bars indicate standard deviations.

them" suggest that the tight integration of an educational system with other classroom activities, including teacher-led discussions, can be beneficial.

Learning of target words was directly measured by the post-test. On average, students answered 89% of cloze exercises correctly, compared to less than 50% in previous studies with the REAP Tutor. A direct comparison to those studies is challenging since the system in this study provided instruction on words that students were also studying as part of their regular coursework, whereas systems in previous studies did not.

9 Discussion and Future Work

We have described a system that enables teachers to find appropriate, authentic texts from the Web for vocabulary and reading practice. A variety of language technologies ranging from text retrieval to POS tagging perform essential functions in the system. The system has been used in two courses by over fifty ESL students.

A number of questions remain. Can language learners effectively and efficiently use such a system to search for reading materials directly, rather than reading what a teacher selects? Students could use the system, but a more polished user interface and further progress on filtering out readings of low text quality is necessary. Is such an approach adaptable to other languages, especially less commonly taught languages for which there are fewer available Web pages? Certainly there are sufficient resources available on the Web in commonly taught languages such as French or Japanese, but extending to other languages with fewer resources might be significantly more challenging. How effective would such a tool be in a first language classroom? Such an approach should be suitable for use in first language classrooms, especially by teachers who need to find supplemental materials for struggling readers. Are there enough high-quality, low-reading level texts for very young readers? From observations made while developing REAP, the proportion of Web pages below fourth grade reading level is small. Finding appropriate materials for beginning readers is a challenge that the REAP developers are actively addressing.

Issues of speed and scale are also important to consider. Complex queries such as the one shown in Figure 3 are not as efficient as boolean queries. The current system takes a few seconds to return results from its database of several million readings. Scaling up to a much larger digital library may require sophisticated distributed processing of queries across multiple disks or multiple servers. However, we maintain that this is an effective approach for providing texts within a particular grade level range or known target word list.

Acknowledgments

This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University; Dept. of Education grant R305G03123; the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420; and a National Science Foundation Graduate Research Fellowship awarded to the first author. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors, and do not necessarily reflect those of the sponsors.

References

Luiz Amaral, Vanessa Metcalf and Detmar Meurers.

2006. Language Awareness through Re-use of NLP Technology. *Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006*.

Jon Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. *Proceedings of InSTIL/ICALL Symposium 2004*. Venice, Italy.

Jon Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. *Proceedings of HLT/EMNLP 2005*. Vancouver, B.C.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology, 56(13). pp. 1448-1462*.

Averil Coxhead. 2000. A New Academic Word List. *TESOL Quarterly, 34(2)*. pp. 213-238.

S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm. 2003. *DOM-based content extraction of HTML documents*. ACM Press, New York.

F. M. Hafiz and Ian Tudor. 1989. Extensive reading and the development of language skills. *ELT Journal 43(1):4-13*. Oxford University Press.

Michael Heilman, Kevyn Collins-Thompson, Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Michael Heilman, Alan Juffs, Maxine Eskenazi. 2007. Choosing Reading Passages for Vocabulary Learning by Topic to Increase Intrinsic Motivation. *Proceedings of the 13th International Conferenced on Artificial Intelligence in Education*. Marina del Rey, CA.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2006. Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. *Proceedings of the Ninth International Conference on Spoken Language Processing*. Pittsburgh, PA.

Thorsten Joachims. 1999. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.)* MIT-Press.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. Draft available at http://www-csli.stanford.edu/∼hinrich/information-retrieval-book.html.

Eleni Miltsakaki and Audrey Troutt. 2007. Read-X: Automatic Evaluation of Reading Difficulty of Web Text. *Proceedings of E-Learn 2007, sponsored by the Association for the Advancement of Computing in Education*. Quebec, Canada.

Matthew Peacock. 1997. The effect of authentic materials on the motivation of EFL learners. *ELT Journal 51(2):144-156*. Oxford University Press.

Kathleen M. Sheehan, Irene Kostin, Yoko Futagi. 2007. SourceFinder: A Construct-Driven Approach for Locating Appropriately Targeted Reading Comprehension Source Texts. *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*. Carnegie Mellon University and International Speech Communication Association (ISCA).