Automatically Generating Reading Comprehension Look-Back Strategy Questions from Expository Texts

Donna M. Gates

Carnegie Mellon University

Final Project for Masters in Computer Assisted Language Learning

Completed May 14, 2008

Automatically Generating Reading Comprehension Look-Back Strategy Questions from Expository Texts

Learning to read is an important skill for both children and adults whether it takes place in their first language or their second language. According to the National Center for Educational Statistics 2003 Assessment (released 2007), 5% of adults (16 years and older) in the US are functionally non-literate. Adults who enroll in literacy programs in the United States desire to improve reading skills for the purposes of acquiring a high-school graduate equivalence degree, improving job skills, reaching career goals, enjoying reading, gaining access to better health care or helping their own children with homework. Resources are limited as literacy programs often rely on volunteer tutors and variable funding sources. Computers are increasingly used to assist in the tutoring process. Improving the computer's ability to aid in this process with useful tools for instructors and students will allow students access to more reading material and provide more opportunities to read. The primary goal for everyone is to increase reading and reading ability.

For instructors, writing questions for reading comprehension exercises is time consuming and difficult. The questions are only useful for a single text. Authoring tools such as Conduit's Dasher (1993), Half-baked Software's Hotpotatoes (2003-2007) and the University of Arizona's MaxAuthor (2007) make the task somewhat easier with tools for creating exercises, and more productive by means of allowing instructors to share exercises on the Internet.

Reading comprehension is composed of several skill levels that work with various language or text chunks: awareness of phonemes (sounds), decoding skills (taking apart words), fluency (ease and speed of decoding), vocabulary knowledge (word meaning), grammar skills (syntax), and strategy skills (meta-learning skills to figure out what the text is about).

The goal of the project described in this report was to create a system that could automatically and accurately generate fact-based reading comprehension questions with expository text for teaching or using a specific look-back reading strategy. The project also aimed to use existing Natural Language programs and knowledge sources as well as implementing new ways to combine their output.

Framework

This project focuses on forming question-answering exercises as part of a reading strategy that will force the student to look back at the text and re-read it. It follows an interactive approach to teaching reading. Interactive teaching means interrupting the user to answer the questions about the text. In both interactive and non-interactive views of teaching reading, answering questions has been a source of assessment as both pre-test and post-test instruments.

Reading is defined as interaction with text for the purpose of understanding it, according to the National Reading Panel Report (National Institute of Child and Human Development [NICD] 2000). To arrive at an understanding of the text is a complex process involving: phonemic awareness, decoding skills, fluency, vocabulary knowledge, grammatical skills, and strategies for understanding text such as finding who, what, where and when events are taking place, why they are happening, drawing inferences and identifying oneself with the text.

Interactionist Theory according to Long, as cited by Chappelle (1998) and Helheimer and Chapelle (2000), claims that language acquisition involves briefly drawing attention to form while focusing on the goal of getting across the meaning. Chapelle's design criteria for CALL systems emphasize interactionist theory such that learners are provided with opportunities to negotiate meaning with the computer. Reading passages and answering questions is an interactive process and when the computer provides feedback, it allows the learner to focus attention on errors and text.

Some researchers do not advocate interactive approaches to teaching reading and they do not advocate interrupting the reader. The Whole-language view posited by Altwerger, Edelsky and Flores (1987) states that reading is a natural language activity that should be based on understanding text for the sake of the information required rather than for teaching how to read. However, Harp (1989) indicates that answering questions is still a useful task within this framework.

Although this project follows the interactive approach to teaching reading, it should be noted that factors such as motivation, interest, and background knowledge all factor into the reader's ability to comprehend a specific text. There is no one right way to improve reading, but all seem to agree that reading more is better.

Background Literature

Question-Answering has been used as a form of assessing reading comprehension for many years (NICH 2000). In the 1980s, reading came to be seen as an interactive process rather than a receptive process. Question answering became one possible reading strategy tool for Dowhower's (1994) three phases of interaction: activating prior knowledge and forcing the

student to learn to self-monitor during the active reading process as well as a tool for recalling information in the post-reading stage. These types of strategies have also been discussed as being beneficial for learning disabled readers (Sorrel 1996).

Reading Strategies

According to the National Reading Panel report (NICHD 2000), strategies are metacognitive procedures that a reader employs for understanding text by verifying what is understood during and after the reading process. Readers become aware of how they are reading.

Raphael (1982) describes the teaching of three categories of Question-Answer Relationships (QARs): *Right There*, *Think and Search* and *On My Own*. The answers to *Right There* questions can be found directly in the reading text.

Garner, Wagner and Smith (1983) reported a study that showed readers as having better performance on answering questions after reading passages when using a look-back strategy in which they went back over the text to find answers. Poor readers did not use the look-back strategies or else they would re-read the entire passage.

Ezell, Kohler, Jarzynka, and Strain (1992) reported that using peer-created QARs improved children's reading comprehension for all three types of questions for low medium and high proficiency readers. Ezell, Hunsicker, Quinque, and Randolph (1996) found that children best were able to retain text-specific reading strategies over a one year period after having been trained in the QAR distinctions.

Multiple choice questions have the drawback of possibly adding information or knowledge rather than purely measuring recall. Whether questions are effective also depends on the level of reading proficiency. Advanced students already possess effective self-monitoring habits and do not require help while reading. Song (1998) showed that EFL students benefited from reading strategy instruction and that less able students benefited more.

Day and Park (2005) describe a taxonomy of comprehension techniques in which questions are an excellent tool for the most basic level of reading comprehension, literal comprehension, in which pertains to understanding explicit facts. Based on current literature, they also claim that these types of questions are best used with low proficiency readers.

Answering questions was found to be one of the effective strategies in teaching adults reading comprehension. (National Institute for Literacy 2005; Day and Park 2005; Dowhower 1994). Questions have been used to activate prior-knowledge, focus attention while reading and re-read text. Low-level readers benefit most from explicit reading strategy instruction (Song 1998) (Dowhower 1994; Garner 1983). Low-level readers have been shown to benefit more from other visual help such as comic strips than do high-level readers (Liu 2004). EFL readers and native Turkish readers benefit from reading strategy instruction according to protocols conducted by Salataci and Akyel (2002).

The National Reading Panel (NICHD 2000) found that questions were a good technique for teaching reading strategies to students (especially with low reading comprehension skills). They also found that strategy instruction may vary with text type. However, the distinction between narrative and expository text comprehension lacks sufficient research. Explicit strategy instruction was recommended as a way to improve comprehension.

Expository Text and Reading

One proven method of instruction recommended by a MATHEMATICA (2006) report is embedded instruction in which the strategy is taught via engaging the student in meaningful text for an independent goal such as understanding a social studies text or a science text. The report also suggests that this sort of expository text is harder than narrative to comprehend because of its unfamiliar styles and unfamiliar topics. Additionally, it is the transition from reading narrative text to expository text that creates new comprehension problems for students around the fourth grade. Little research has been done to distinguish the specific strategy needs for these two types of text (National Reading Panel [NRP] 2000). The NRP also complained that many studies did not make it explicit as to which type of text was used in their research findings.

The NRP also reported that computer technology shows promise in the are of reading comprehension especially with respect to speech recognition and multi-media. It indicated that there is much interest in this area but very little in terms of educational research studies.

CALL

Computer Assisted Language Learning (CALL), according to Chapelle (1998) should support interaction between the learner and the computer. The computer should act as a participant in the interaction. Learners should be able to focus on salient linguistic features of the task. For reading comprehension, this can take the form of traditional teacher defined question answering exercises of even automatic cloze question fill in the blank exercises (Brown, Frishkoff and Eskenazi 2005).

Wilson (1997) indicates two problems with courseware materials. First they are often hard to fit to the different abilities of various students. Secondly, it is difficult to create exercises that

can be re-generated with new examples for students when reviewing the same material. To solve these problems, she automatically generated cloze exercises from corpora using simple search techniques over tagged corpora.

With the increased use of computers to aid in language learning, automated question answering is simple, easy to score, and infinitely patient. It frees up instructors for complex communication tasks. Hegelheimer and Chapelle (2000) state that interactive approaches to reading via the computer also assist in tracking the student's activities of noticing and can provide a variety of glosses for vocabulary acquisition. Anderson (2003) reported a study in which ESL and EFL students were interviewed to determine common reading strategies used with on-line text. These included clicking and re-reading text for better understanding.

Mostow et al. (2004) and Brown et al. (2005) used automatically generated cloze multiple-choice vocabulary exercises to assess vocabulary knowledge. Brown et al (2005) showed that the automated questions created to go with the REAP (Heilman, Collins-Thompson, Callan and Eskenazi 2006) reading system were comparable in assessment to hand-written vocabulary assessment questions. REAP uses text classification techniques to mine texts from the web and determines the reading level based on the vocabulary in the texts. Mostow et al.'s (2004) results for the Project Listen system were mixed based on reading levels for questions given during the reading of the story versus in the post-test phase showing differing improvements in responses to questions for grades 1-3 (during the story) versus 4-6 (in the post-test). Research on the automatic assessment of reading comprehension (Roturo and Litman 2005) has used techniques from question–answering research for finding answers to user-generated questions on the web. Questions were created by instructor or taken from a standardized test.

Motivated by the recommendation that strategy teaching with fact-based questions would benefit students, Beck, Mostow and Bey (2003) tested students with fact-based questions and showed that such questions did not indicate improvement for readers. However, these questions were generic in form and meaning and not taken directly from the texts.

Harben (1999) describes a system for listening and multi-media interaction in which students were given exercises (including multiple-choice questions) and allowed to re-view or re-listen to material. The students' actions (mouse clicks, scrolls, re-viewings, and re-plays) were recorded on the computer for later analysis. Students were given a questionnaire at the end of the exercises to determine their perception of the materials, exercises as well as the usefulness of the re-viewing/re-reading/re-listening capabilities of the system. A large majority of the students reported that it was "useful" to review the materials before answering the questions.

Language Technology and Question Answering

Wolfe (1977) created a program to generate questions from text using simple pattern matching techniques available at the time. He further hypothesized that the incorporation of NLP tools would eventually make this a more accurate task. The advent of FrameNet (Fillmore and Baker 2001) and PropBank (Palmer 2005) dictionaries, supporting programs for identifying semantic roles such as ASSERT (Pradhan et al. 2005) , as well as easily available high quality parsers such as Stanford's NL parser and Connexor's syntactic parser, and large semantically organized dictionaries such as WordNet (Fellbaum 1998) have made this possible. BBN's Identifinder (Bikel, Schwartz, & Weischedel 1999) is a system for annotating corpora with named entities and useful semantic classes such as PERSON, ORGANIZATION, and TIMEX to name a few. Connexor's Machinese Semantics system also labels named entities and semantic classes of nouns.

Mitkov and Ha (2003) and Chen Yiou and Chang (2006) developed systems to create WH multiple choice and cloze multiple choice questions respectively. They used NLP tools including word net, semantic knowledge bases and parsing. The latter implemented some heuristic patterns to filter text.

Q&A research (e.g., *Guru* as described in Prager, Brown , Coden and Radev 2000) is directed at obtaining answers from large sets of documents such as the web in order retrieve documents that answer questions posted by the user such as "Who is the leader of France?". For this project, I hope to use transformation rules to produce questions from annotated text as well. Deep Read (Hirschman, Light, Breck and Burger (1999) is an automated system using various linguistic modules for finding answers in single sentences for Q&A. The answers to reading comprehension questions generated from a passage can be found in a single sentence in a known text so that complicated modules will not be necessary for finding the answer. The jeopardy model (Wang, Smith and Mitamura 2007) for Q&A also takes statements and transforms them into questions for later use by the Q&A system. The text chosen for the project proposed here has been used in Q&A testing by using reading comprehension style human created questions and answers to evaluate Q&A systems (Rotaru and Litman 2005).

Synthesis

Questions have been shown to help low-proficiency readers comprehend text as both pre-reading and concurrent-reading strategies especially when taught as a re-reading or look-back strategy explicitly with other reading strategies. CALL has already shown the capacity to generate cloze questions for assessment of reading comprehension. Recent advances in NLP and semantic annotations sparked by Q&A research now make it possible to annotate texts with sufficient information that can be used for other applications.

If factual questions can be generated automatically, then CALL can also improve expository text reading comprehension for low-proficiency readers by providing not only unlimited practice, but increase the availability of level-specific reading exercises from material of interest to students.

Project Description

The goal of this research is to discover if text-based factual reading comprehension questions can be accurately automatically generated using existing NLP tools and hand-written transformation rules. Can an accuracy rate of at least 80% grammatical and semantically coherent questions be attained? This project combines existing NL programs, existing knowledge sources and newly created knowledge in a unique way to automatically produce questions and answers for reading strategy practice.

Definitions

Reading comprehension for the purposes of this project will be defined as the information or facts that the participant/reader understands and retains from reading an expository text.

Reading strategy in this project is the presentation of *Right-there-in-the-text* type questions shown below a reading passage to promote using a look-back strategy (i.e., re-reading the text to find the clickable answers encoded in the text).

A reading comprehension question is a WH-fact based question directly derived from an expository text**.** Reading comprehension question accuracy will be measured based on the grammaticality, semantic correctness and practicality of the questions produced from the text.

The questions must adhere to grammatical standards of syntax and semantics. The questions must also have answers that can be found in the text of the passage.

Corpora

The Canadian Broadcast Company produces English news reports for children age 9-12. Texts from 1999 and 2000 were collected by MITRE Corporation and used in TREC evaluations (Leidner 2003 and Dalmas 2003). There are 52 test files and 73 training files. I used the training set for development and the test set for evaluating the question production of my program. The latter set remained unseen until the program was frozen for evaluation.

System Overview

Figure 1 illustrates the question and answer system. Boxes in the diagram represent programs and ovals represent knowledge sources. Output is represented by text only. Broken lines represent components and knowledge sources that were added after the initial 2007 prototype (solid lines) was constructed.

The input is a CBC4Kids text passage. Since CBC passages are already segmented into sentences, no segmenter was used in this version of the system. Each sentence is parsed by the Stanford syntactic parser to obtain a parse tree with surface forms of words and to determine the stem or root of each word. The BBN Indentifinder named entity annotator labels each sentence with proper name types such as PERSON, ORGANIZATION, TIMEX. Where possible, the ASSERT semantic role labeler annotates each sentence with PropBank (Palmer 2005) argument labels such as ARG0, ARG1, ARGTMP, and PASSIVEARG0.

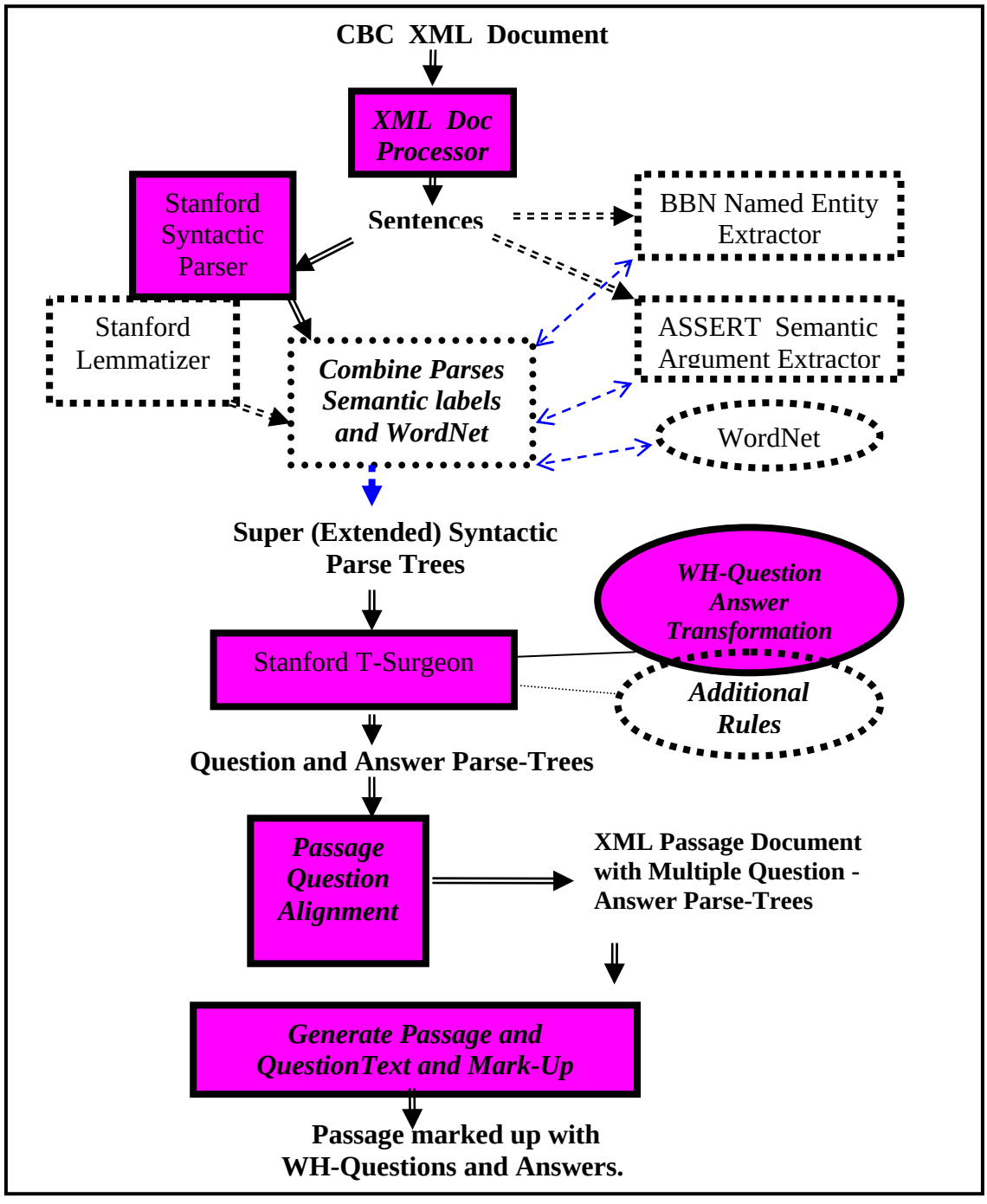


Figure 1 System Diagram showing data flow and components

Common nouns from the parse tree are looked up in a pre-sorted WordNet derived database to obtain general semantic class information similar to semantic class features extracted in Yoo et al (2008) such as animacy (person). Words in English WordNet are organized into sets of synonyms (synsets) and arranged hierarchically with respect to other synsets: woman/lady is a female which is a person/human being, which is a mammal, which is etc.

Animacy (human) corresponds to words that fall under the hierarchical synsets in WordNet for **person** (*girl*), **organization** (*company*), and **region** (*city*) while temporal phrases use **period** (*today*), **time_unit** (*month*). These WordNet features were determined by the author to be useful in determining the type of WH-phrase to generate (e.g., *who/whom* = person, *what* = non-human, *when*= temporal).

These three pieces of semantic knowledge along with the syntactic parses are combined to form a super parse tree (as shown on the left in **Figure 3**) for each sentence.

T-Surgeon Transformations

Each super parse tree is passed to Stanford NLP's T-Surgeon (Levy and Andrew 2006). Tsurgeon requires a set of user-defined hand written rules to perform transformations on trees. The rules are hand-written by the author and designed to transform declarative sentence trees into WH-question trees and to create matching answer trees.

The rules are used to transform parse trees into output trees if they match the regular expression portion of the T-Surgeon rule. A question's corresponding answer tree must be produced in order to mark the boundaries of the phrase that correspond to the correct answer within the sentence. The regular expression portion of the rules is the same for the question and the answer. The following example is a simplified version of the rule designed to target subject NPs with noun heads.

Filename: whnpsubj_vp_question

```
ROOT=root < (S=s [< ((NP=subj < /^NN/) $+ (VP=vp < (VBP|VB|VBZ|VBD|MD)))] [!< S]  <  (/^\./=period <
/^\./=period2))              ;;  find the highest subject NP followed by a tensed VP

insert (RULE WHNPSUBJ_VP)  $+ s    ;;  add this node to the left of the S node
relabel s QUESTION                 ;;  relabel the S node as QUESTION (marks it for later processing)
insert (WHNP who_or_what) $+ subj  ;;  insert a WHNP to the left of the current subject.
delete subj                        ;; delete the subject
relabel period QMARK               ;; change the end period in the tree to a question mark.
```

The first part of the rule consists of the tregex (regular expression) pattern that matches against the parse tree. Nodes with labels (e.g., *=subj*) can be used in the transformations. The rest of the lines are the instructions for moving, deleting, renaming, and inserting elements in the tree in order to transform it into the target tree. The ";" comments in the rule above are for illustration purposes. The above rule (along with an extra transformation rule that removes parenthetical expressions) will transform the following parse tree:

The school has turned its one-time metal shop - lost to budget cuts almost two years ago - into a money-making professional fitness club.

```
(ROOT
  (S
    (NP (DT The) (NN school))
    (VP (VBZ has)
      (VP (VBN turned) (NP (PRP$ its) (JJ one-time) (NN metal) (NN shop))
        (PRN (: -)
          (S
            (VP (VBN lost)
              (PP (TO to) (NP (NN budget) (NNS cuts)))
              (ADVP
                (NP   (QP (RB almost) (CD two)) (NNS years))
                (RB ago))))   (: -))
        (PP (IN into)
          (NP (DT a) (JJ money-making) (JJ professional) (NN fitness) (NN club)))))
    (. .)))
```

Into:

Who or what has turned its one-time metal shop into a money-making professional fitness club?

```
 (ROOT (RULE WHNPSUBJ_VP)
  (QUESTION (WHNP who_or_what)
    (VP (VBZ has)
      (VP (VBN turned)
        (NP (PRP$ its) (JJ one-time) (NN metal) (NN shop))
        (PP (IN into)
          (NP (DT a) (JJ money-making) (JJ professional) (NN fitness) (NN club)))))
    (QMARK QMARK))
```

Currently, there are 14 Tsurgeon rules for transforming sentences into WH phrases and questions: subject rules, direct object rules, passive agent rules, and temporal expression rules (basic question rules). In addition, there are rules for transforming the question trees into the correct formats (clean-up rules), for simplifying/shortening the question by removing relative clauses and PPs where possible (simplifying rules), and for refining the WH words by using semantic information (animacy rules). The rules must be combined in a specific order relative to the general question rule being applied. Figure 2 shows an example of the rules applied to sentences containing passive agents.

whnpbyagent_aux_question :: create the basic wh question for passive agents

- remove_trace_relative :: remove relative clause from the trace PP's NP
- inanimatepp_question :: use WHAT if TRACE NP is an instrument
- animatepp_question :: use WHOM if animacy is human
- remove_trace_pp :: remove the trace PP (cleanup)
- pp-comma-cleanup_question :: remove fronted PPs from the sentence
- cccleanup_question :: remove sentential conjunctions
- advp-cleanup_question :: remove some adverbs
- ppcleanup_question :: remove extra PPs from the sentence
- punctuate_question :: fix the end punctuation
- remove_nqs ::remove the tree if not flagged as a question
- remove_prn :: remove the parentheticals
- move_sentential_obj :: move a fronted sentential adjunct to after the verb

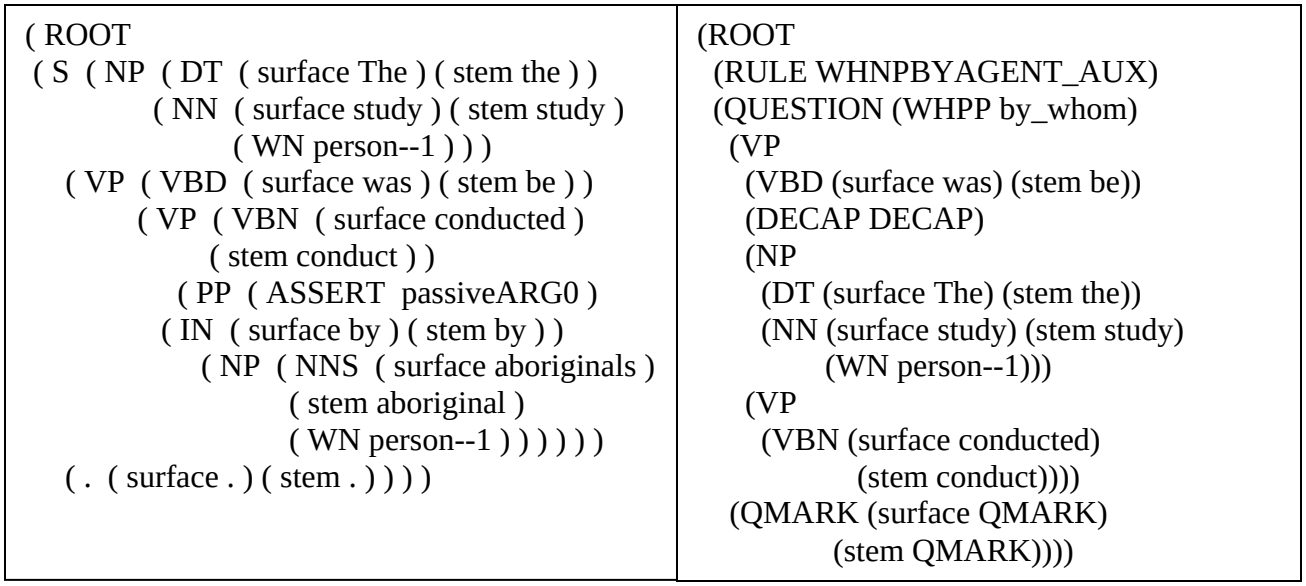
Corresponding ANSWER rules

whnpbyagent_aux_answer :: create tree with start/end phrase markers around answer

- remove_nas :: remove the tree if not flagged as an answer

Figure 2 Example rule ordering for passive agent WH-question-answer pairs.

Figure 3 shows an example of a super parse tree (left) that is the input sent to Stanford's Tsurgeon program. Tsurgeon used the transformation rules shown above to produce the tree on the right for the WH question: *By whom was the study conducted?*



```
( ROOT
 ( S  ( NP  ( DT  ( surface The ) ( stem the ) )
            ( NN  ( surface study ) ( stem study )
                  ( WN person--1 ) ) )
   ( VP  ( VBD  ( surface was ) ( stem be ) )
         ( VP  ( VBN  ( surface conducted )
                      ( stem conduct ) )
               ( PP  ( ASSERT  passiveARG0 )
                     ( IN  ( surface by ) ( stem by ) )
                     ( NP  ( NNS  ( surface aboriginals )
                                  ( stem aboriginal )
                                  ( WN person--1 ) ) ) ) ) )
   ( .  ( surface . ) ( stem . ) ) ) )
```

```
(ROOT
  (RULE WHNPBYAGENT_AUX)
  (QUESTION (WHPP by_whom)
    (VP
      (VBD (surface was) (stem be))
      (DECAP DECAP)
      (NP
        (DT (surface The) (stem the))
        (NN (surface study) (stem study)
            (WN person--1)))
      (VP
        (VBN (surface conducted)
             (stem conduct))))
    (QMARK (surface QMARK)
           (stem QMARK))))
```

Figure 3 Super parse tree and Output question tree

The question trees are then used to create question strings by way of a simple generation script written in Perl that will simply find the surface terminal nodes in the tree and format them into a properly spaced and punctuated English text string. This script also makes sure the finite verbs are properly formed for subject WH-phrases (*Who/What is…*) and that the verb DO is properly conjugated when added to the tree for question transformations requiring DO-support (e.g., *When does the mayor want to leave?* produced from *The mayor wanted to leave next week.*)

In addition, 14 answer transformation rules convert the same input tree into an answer tree. These rules contain the same initial pattern as its respective question rule and adds beginning and ending tags to the tree to indicate the answer. This is then generated with a Perl

script into an English string containing xml tags marking the span of the answer phrase within the sentence: *The mayor wanted to leave<answert>next week</answer>*.

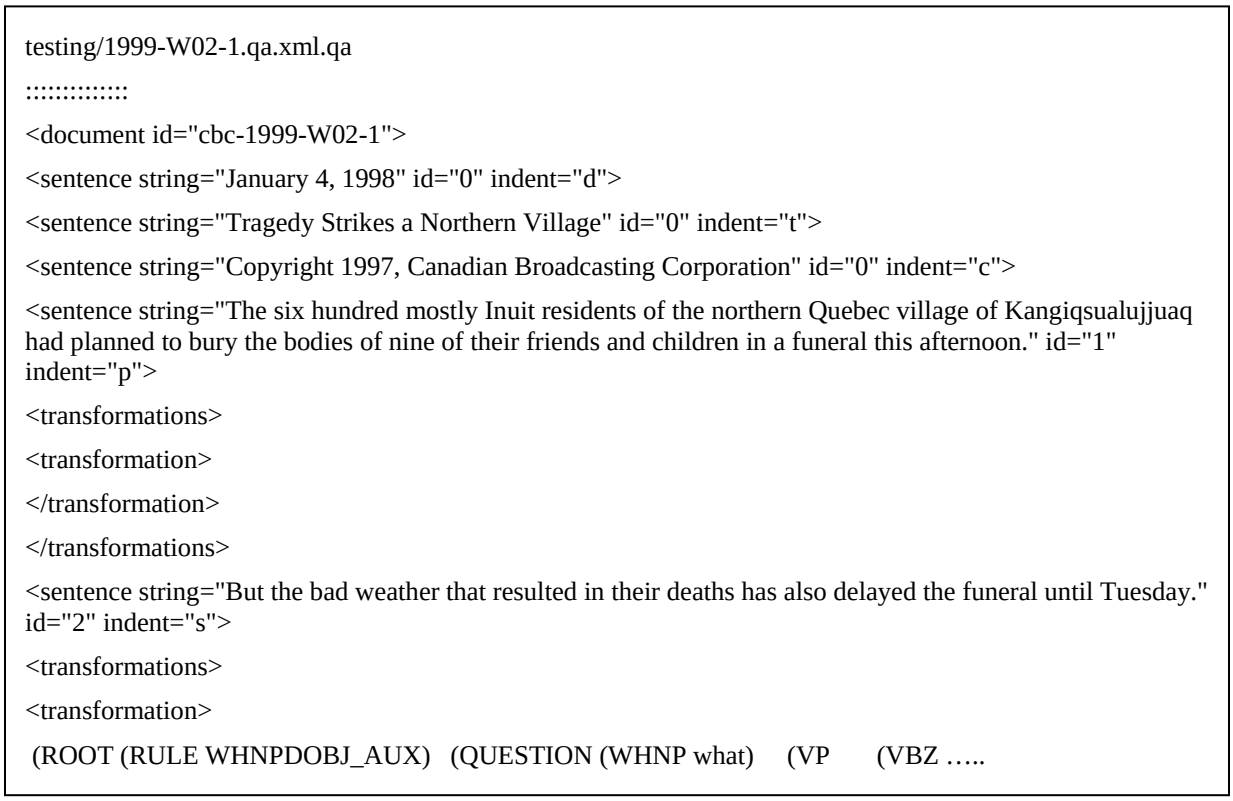


Figure 4 XML marked-up sentences and stored trees for questions and answers.

The set of all questions and answers for each passage are stored together with their xml marked-up sentence which also contains formatting information (title, copyright and paragraph breaks) from the original CBC xml document (Figure 4). The final HTML output is produced by choosing among the available questions automatically. These are then converted into HTML format for display purposes. Each time the HTML files are re-generated, the questions and answers may vary due to random selection. At present, the user can not re-generate the HTML files.

WH Question Formation and Selection

Since the goal of this project is to automatically generate fact-based questions from text, the types of questions must be restricted to those that can be reasonably generated using T-Surgeon transformation rules. The following syntactic and semantic restrictions apply to candidate sentences for question formation:

- If there is a pronoun, it must have a reasonably certain antecedent NP (equal in person/gender/number) nearby.
- Temporal expressions are used if they have anchored or anchorable (e.g., *then* will not be used in a WHEN question since it is not easy to anchor nor should it be anchored in all cases).
- Phrases that answer *WHO, WHAT, WHEN*[1] type questions are used; *HOW* and *WHY* type phrases will not be used.
- The following syntactic functions are used as candidates for being replaced with WH-phrases Syntactic restrictions: *SUBJ, OBJ, PASSIVE by PPs*, some sentence level argument or adjunct PPs such as dates, and times.
- Named entities and semantic roles are only used if they are reliably extractable and categorized.

In the following example, the question conforms to these guidelines. The question is created structurally from the sentence using the date in the example below.

Sentence*: The Inuit children were killed by a storm on April 10.*

Question: *When were the Inuit children killed by a storm?*

Additionally, questions formed from complex sentences are simplified during transformation. For example, in the following sentence,

the Inuit children killed in the winter storm will be mourned by the entire community,

the subject contains a reduced relative clause which is removed from the final question tree prior to question generation to produce the simplified text: *By whom will the Inuit children be mourned?*

[1] WHERE and HOW MANY type questions will be added in the future.

Several possible questions may be generated from this sentence:

Who will mourn the Inuit children?

By whom will the Inuit children be mourned?

Who will be mourned by the entire community?

For display purposes, at most one question is randomly selected for a given sentence or answer. Multiple questions-answers for a sentence are valuable for generating question-answer variations for the same passage when a student re-reads the same passage later.

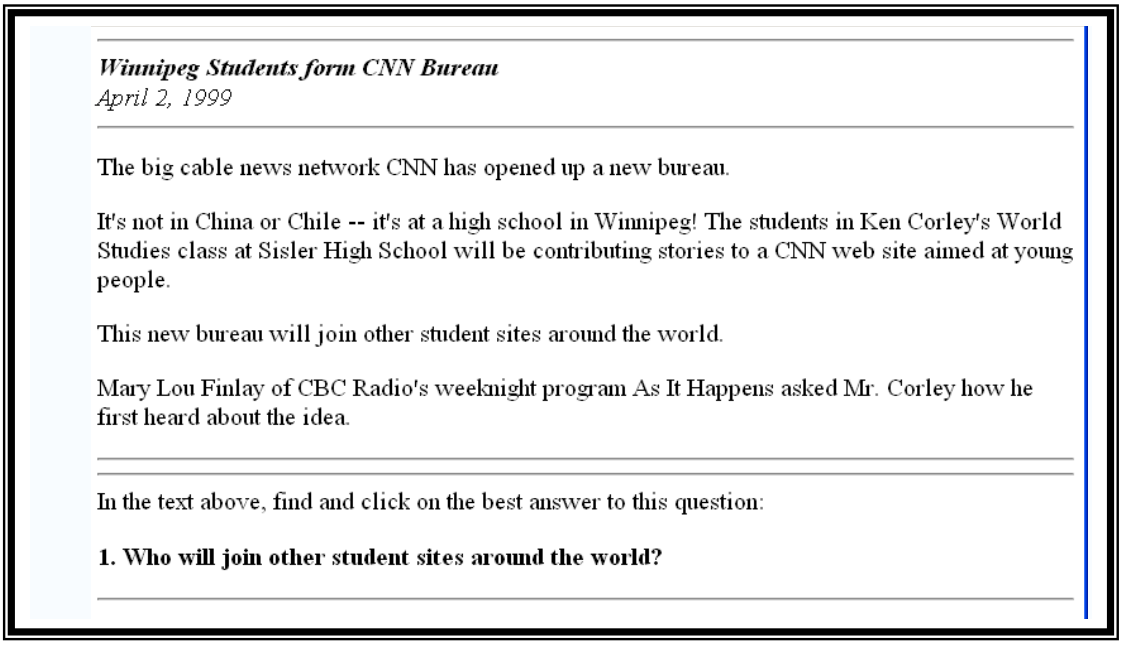
Current question-answering and testing theory for reading recommends that an exercise contain no more than five questions for a given reading comprehension passage (K. Koda personal communication, March, 2007). Therefore, each reading passage exercise displays at most five questions depending on passage length. The question-answers are randomly selected from the sentences such that no sentence is used more than once as an answer to a question.

Final Output

For the purposes of this project, reading comprehension is defined as the understanding and retention of facts found in expository text. Thus, the final output of the program consists of a reading passage along with questions. Questions appear below their relevant passage in order to promote look-back re-reading rather than a pre-reading strategy. The same passage will be displayed with appropriate clickable regions for each subsequent question, but only one question will appear at a time. Figure 5 illustrates a sample reading passage with a sample question. Clicking on the correct answer in the passage will give a *Correct!* feedback response. Clicking on the answer sentence outside the correct answer gives a *Please be more specific* feedback response. Lastly, clicking outside of the answer sentence will elicit a *Please look elsewhere* feedback response. These clickable areas along with pop-up responses are illustrated in appendix

A's screen image. The user may move to the next question at will. Currently, the system does not maintain a record of responses.

Links to recently generated HTML files containing reading passages with questions can be found at the following web page: http://www.cs.cmu.edu/~dmg/MCALL/mcall.html.



Winnipeg Students form CNN Bureau
April 2, 1999

The big cable news network CNN has opened up a new bureau.

It's not in China or Chile -- it's at a high school in Winnipeg! The students in Ken Corley's World Studies class at Sisler High School will be contributing stories to a CNN web site aimed at young people.

This new bureau will join other student sites around the world.

Mary Lou Finlay of CBC Radio's weeknight program As It Happens asked Mr. Corley how he first heard about the idea.

In the text above, find and click on the best answer to this question:

1. Who will join other student sites around the world?

Figure 5 Sample Passage with Question.

Evaluation

In order to evaluate the system, the accuracy of the question-answer pairs produced by the system was measured. Accuracy in this case refers to both the grammaticality of the questions and the practical usefulness of the questions given the text and the answers. In order to accomplish this, each question was scored as to whether it was syntactically grammatical and whether it made sense semantically in the context of the text. The answer must be easily found. Additionally, the system requires that there is one clearly correct answer in the text since the HTML question answer key does not allow other possibly valid answer phrases to be selected. The correct answer must be selected in the sentence from which the question was generated.

The goal of this project was to achieve a minimum of 80% accuracy on evaluation data consisting of the 52 passages from the CBC4Kids evaluation set. These were not used for development of the system. The questions were rated by hand by the author and by a second independent grader. Each question was given one of the following grades:

- ***P***erfect, grammatical well-formed question with an obvious answer.
  - *Who conducted a study?*
  - *(*Source*: Aboriginals conducted a study.)*
- O***K***, grammatical, possibly awkward question with an obvious answer.
  - *Who or what conducted a study?*
  - *(*Source*: Aboriginals conducted a study.)*
- ***B***ad, very awkward question, possibly incorrect WH phrase (*what* vs *who*) but the answer is still obvious.
  - *What conducted a study?*
  - *(*Source*: Aboriginals conducted a study.)*
- ***F***ailure, ungrammatical, not well understood question and/or impossible to find answer.
  - *When $180 did the government pay the people a rebate?*

A question is *acceptable* if it is *perfect* or *ok*. The author/system developer judged the output to be 85% *acceptable* while the independent grader judged the output to be 81% *acceptable*. The scores are broken down and shown in **Table 1**. These are based on the independent grader's scoring. The table shows that the subject rules produced the greatest number of question-answer (QA) pairs. The temporal expression rules were more restricted and produced fewer questions but produced QA pairs with the highest accuracy.

Wh-Question Transformation	Quantity of type	Perfect	Ok	Bad	Failure
Subject Wh-Phrase					
NP Subject *Who conducted a study?* *The aboriginals conducted a study.*	444	80%	6%	13%	2%
Gerund Subject *What is a new experience?* *Conducting research is a new experience.*	0	0%	0%	0%	0%
Wh-phrase from Direct Object under Verb Phrase					
Do-support: *What did the aboriginals conduct?* *The aboriginals conducted a study*	119	55%	8%	25%	12%
Auxiliary verb *What were the aboriginals studying?* *The aboriginals were studying the effects of alcohol.*	52	58%	10%	25%	8%
Modal verb *What will the children build next?* *The children will build a new playground next.*	24	63%	13%	17%	8%
Passive Agent *By whom was the first study completed.* *The first study was completed by the company last year.*	11	91%	0%	9%	0%
Wh-phrase (When) from Temporal Expressions ***When was the study conducted?***					
From PP under S node ***The study was conducted in May.***	19	89%	5%	5%	0%
From NP under S node ***The study was conducted last May.***	1	100%	0%	0%	0%
From PP under VP node ***In May the study was conducted.***	14	86%	7%	0%	7%
From NP under VP node ***Last May the study was conducted.***	9	100%	0%	0%	0%
ALL Transformations	693	75%	6%	15%	4%

Table 1 Distribution of question transformation types and evaluation scores on unseen data.

NLP Errors and Problems

Errors in question formation were due to several factors: parsing errors, question pattern errors, WordNet ambiguity and BBN annotation errors or missing annotations. Lexical ambiguity from WordNet resulted in some words being miscategorized as human nouns as with *dish (plate* vs *attractive woman),* or *mind (thinking organ* vs *smart person)* such that the WH question was generated with *WHO* instead of *WHAT*. Lexical disambiguation or meaning frequency constraints will be required for better accuracy. BBN, in some cases, misidentified a proper name as a PERSON such as *Cassini* (a spacecraft). This system can be retrained with additional data or a better annotator might be used in its place. Since the parser did not distinguish which of multiple nouns in an NP was the head of the phrase *("contact person"* = **(NP (NN contact) (NN person))**), the transformation patterns were not able to identify the exact noun animacy for determining WHO vs WHAT in some cases. In the future, it will be necessary to identify the head word in the NP prior to determining animacy.

Discussion

Issues and Concerns

This project emphasizes one component of one type of reading strategy. The National Reading Panel and QAR researchers have stated that reading strategies are best taught together. However, the goal of this research is to determine whether it is possible to automatically generate fact-based questions accurately as a prompt for using a look-back strategy. This does not preclude the inclusion of such a component in a more extensive strategy instruction approach.

The question-answer generation relies heavily on accurate output from the parser, named-entity extractor and semantic role labelers. Because of this, programs (when freely available) with better capabilities in these areas need to be incorporated into the system's architecture to improve the semantic constraints used by the transformation rules.

Results

The project described here was intended to demonstrate the capability of current NLP tools and hand-made transformations to generate reading comprehension questions with at least 80% grammatical precision and semantic accuracy for children's expository reading passages from authentic text. The system produced acceptable questions 81% of the time for texts that had not been used in system development. While this is not ideal, it is at least a promising result.

Pedagogical Implications

If these types of questions can be generated automatically, they could be used for computer aided tutoring of reading. This, in turn, would save time and effort for instructors and

allow students to seek more practice and help with on-line texts of interest. Due to time constraints, the design and implementation of an experiment to study the effects of using this system on student reading comprehension or strategy use was beyond the scope of this project. Such an experiment required the completion of the proposed system. This will be considered as future work once the question generation system is made more reliable.

Conclusion and Future Work

While it is possible with a few simple NLP tools to develop fact-based WH-questions for use in reading practice, the process is not yet completely automatic. With acceptable generation achieving a score of 81%, human input is still required to weed out ungrammatical questions and pragmatically poor questions. However, ungrammatical questions may eventually be reduced significantly by better use of semantic resources.

In its current form, the project described here would be useful as part of an authoring tool to assist instructors in creating new reading passage question-answer exercises. Once the question-answer output can be made reliable, self-paced readers would have access to new reading passages that may be of more interest to them along with automatically generated questions.

Acknowledgements

The material described in this paper is based upon work completed for program of study at Carnegie Mellon University and was partially supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

I wish to give particular thanks to Eric Riebling for his help with accessing and running the BBN Identifinder and ASSERT programs. I also wish to thank Sarah Wells for her patience and understanding, and her help with the evaluation. I thank Dr. Maxine Ezkenazi and Dr. Chris Jones for the opportunity to participate in the MCALL program. I owe a special thank you to Dr. Lori Levin for many years of supporting me and advising me on this project.

I also wish to thank my mother for her love and sacrifices and for the promise I made to her.

References

Altwerger, B., Edelsky, C. & Flores, B. M. (1987). Whole Language: What's new? *The Reading Teacher,* 41, 145-154.

Anderson, N. (2003). Scrolling, Clicking, and Reading English: Online reading strategies in a second/foreign language. *The Reading Matrix.* 3 (3).

Beck, J. E., Mostow, J., & Bey, J. (2004). Can automated questions scaffold children's reading comprehension? *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Maceio, Brazil.

Bikel, D., R. Schwartz, and R. Weischedel. (1999). An Algorithm that Learns What's in a Name. *Machine Learning—Special Issue on NL Learning*, 34, 1–3.

Brantmeier, C. (2004). Statistical procedures for research on L2 reading comprehension: An examination of ANOVA and regression models. *Reading in a Foreign Language,* 16 (2).

Brown, J., Frishkoff, G., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of HLT/EMNLP 2005*. Vancouver, B.C.

Chapelle, C.A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, 2 (1), 22-34.

Conduit. (1993). Dasher [computer software]. University of Iowa, Iowa City, IA.

Connexor Oy Inc. (2007) Machinese Syntax and Machinese Semantics [computer software]. Helsinki Business and Science Park, Finland: http://www.connexor.com/.

Dalmas, T., Leidner, J.L., Webber, B., Grover, C. & Bos, J. (2003). Generating Annotated Corpora for Reading Comprehension and Question Answering Evaluation. In *EACL-2003 workskhop on Natural Language Processing (NLP) for Question-Answering.* Budapest, Hungary.

Day, R., & Park, J. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, 17 (1).

Dowhower, S. L. (1994). Repeated reading revisited: Research into practice. *Reading & Writing Quarterly: Overcoming Learning Difficulties*. 10(4), 343-358.

Ezell, H. K., Hunsicker, S. A., Quinque, M. M., & Randolph, E. (1996). Maintenance and generalization of QAR reading comprehension strategies. *Reading Research and Instruction*, 36 (1), 64-81.

Ezell, H. K., Kohler, F. W., Jarzynka, M., & Strain, P. S. (1992). Use of peer-assisted procedures to teach QAR reading comprehension strategies to third-grade children. *Education and Treatment of Children*, 15 (3), 205-27.

Fellbaum, C. (1998), ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Ma.

Fillmore, C. and Baker, C. (2001): Frame Semantics for Text Understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, NAACL, Pittsburgh, Pa.

Garner, R., Wagner, S., & Smith, T. (1983). Externalizing question-answering strategies of good and poor comprehenders. *Reading Research Quarterly*, 18(4), 439-447.

Graham, L. & Wong, B.Y.L. (1993). Comparing two modes of teaching a question-answering strategy for enhancing reading comprehension: Didactic and self-instructional training. *Journal of Learning Disabilities*, 26 (4), 270-279.

Half-Baked Software Inc. (2003-2007) Hotpotatoes [computer software]. Victoria, B.C.: http://www.halfbakedsoftware.com.

Harp, B. (1989). When the Principal Asks: "Why don't you ask questions?" *The Reading Teacher* , 42(8), 638-639.

Harben, P. (1999). An exercise is applying pedagogical principles to multimedia CALL materials design. *ReCALL*, 11(5), 25-33.

Hegelheimer, V. & Chapelle, C.A. (2000). Methodological issues in research on learner-computer interactions in CALL. *Language Learning & Technology*, 4(1), 41-59.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2006). Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the Ninth International Conference on Spoken Language Processing.*

Hirschman, L., M. Light, E. Breck, and J. D. Burger. (1999). Deep read: A reading comprehension system. In *Proceedings, ACL'99*, 325-332, University of Maryland.

Huang, H. & Liou, H., (2007). Vocabulary learning in an automated graded reading program. *Language Learning & Technology.* 11(3), 64-82.

Kipper, K., Palmer, M., & Rambow,O. (2002). Extending PropBank with VerbNet Semantic Predicates. *Workshop on Applied Interlinguas, held in conjunction with AMTA-2002.* Tiburon, CA, October.

Klein D. & Manning, C. (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002),* 3-10, Cambridge, MA.

Learning Point Associates (2005). The effects of technology on reading performance in the middle-school grades: A meta-analysis with recommendations for policy. Retrieved April 2, 2007 from http:/www.ncrel.org/tech/reading/index.html. Naperville, IL: P. D. Pearson, R.E. Ferdig, R.L Blomeyer Jr., & J. Moran.

Leidner, J.L., Dalmas, T., Webber,.B., Bos,J., & Grover,C. (2003). Automatic Corpus Layer Annotation for Evaluating Question Answering methods: CBC4Kids. *EACL-2003 Workshop for Linguistically Interpreted Corpora (LINC-03)* Budapest, Hungary.

Levy R. and Galen Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *Proceedings of LREC 2006*.

Liu, J. (2004). Effects of comic strips on L2 learners' reading comprehension. *TESOL Quarterly*, 38 (2), 225-243.

Mackey, A. (1999). Input, interaction and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, 21(4), 557-587.

MATHEMATICA Policy Research Inc. (2006). *The National Evaluation of Reading Comprehension Interventions: Design report* (MPR Reference No. ED01CO0039/0010). Retrieved March 10, 2007, from http://www.mathematica-mpr.com/publications/PDFs/readcomp.pdf

Mitkov, R. & Ha, L.A. (2003). Computer-aided generation of multiple-choice tests. In Proceedings *of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, 17-22, Edmonton, Canada.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., & Valeri, J. (2004). Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning*, 2, 97-113.

Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Lernaing & Technology*. 11(3), 107-129.

National Center for Educational Statistics. (2007). *Literacy in everyday life: Results from the 2003 assessment of adult literacy* (US Department of Education NCES 2007-480). Washington, D.C.: M. Kutner, E. Greenberg, Y. Jin, B. Boyle, Y. Hsu, & E. Dunleavy. Retrieved May 10, 2007, from http://nces.ed.gov/pubsearch /pubsinfo.asp?pubid=2007480.

National Institute of Child and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implication for reading instruction: Reports of the sub-groups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31 (1), 71-106.

Phakiti, A. (2003). A closer look at gender and strategy use in L2 reading. *Language Learning,* 53(4), 649-702.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H. & Jurafsky, D. (2005). Support Vector Learning for Semantic Argument Classification. *Machine Learning*, 60(1), 11-39.

Prager, J., Brown, E., Coden, A. & Radev, D. (2000). Question-Answering by Predictive Annotation, *Proceedings of SIGIR* 2000, Athens, Greece.

Raphael, T. E. (1982). Question-answering strategies for children, *Reading Teacher*, 36 (2), 186-190.

Raphael, T. E. & McKinney, J. (1983). An examination of fifth- and eighth-grade children's question-answering behavior: An instructional study in metacognition. *Journal of Reading Behavior*, 15 (3), 67-86.

Reynar.J. and Adwait Ratnaparkhi, A.(1997). A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington, D.C.

Rotaru, M & Diane J. Litman, D. J. (2005). Improving question answering for reading comprehension tests by combining multiple systems. In *Procedings of American Association for Artificial Intelligence 2005*.

Salataci, R. & Akyel, A. (2002). Possible effects of strategy instruction on L1 and L2 reading. *Reading in a Foreign Language*, 14(1).

Song, M., (1998). Teaching reading strategies in an ongoing EFL university reading classroom. *Asian Journal of English Language Teaching*, 8, 41-54.

Sorrell, A. L. (1996). Triadic Approach to reading comprehension strategy instruction. Presented at the *32nd Annual Conference of the Learning Disabilities Association of Texas.* October 25, Austin, TX.

University of Arizona. (2007). MaxAuthor [software]. Tucson, AZ: Computer Aided Language Instruction Group, http://cali.arizona.edu/docs/wmaxa.

Wang, M., Smith, N. & Mitamura, T. (2007) What is the Jeopardy Model? A Quasi-Synchronous Grammar for Question Answering, *In Proceedings of EMNLP '07*.

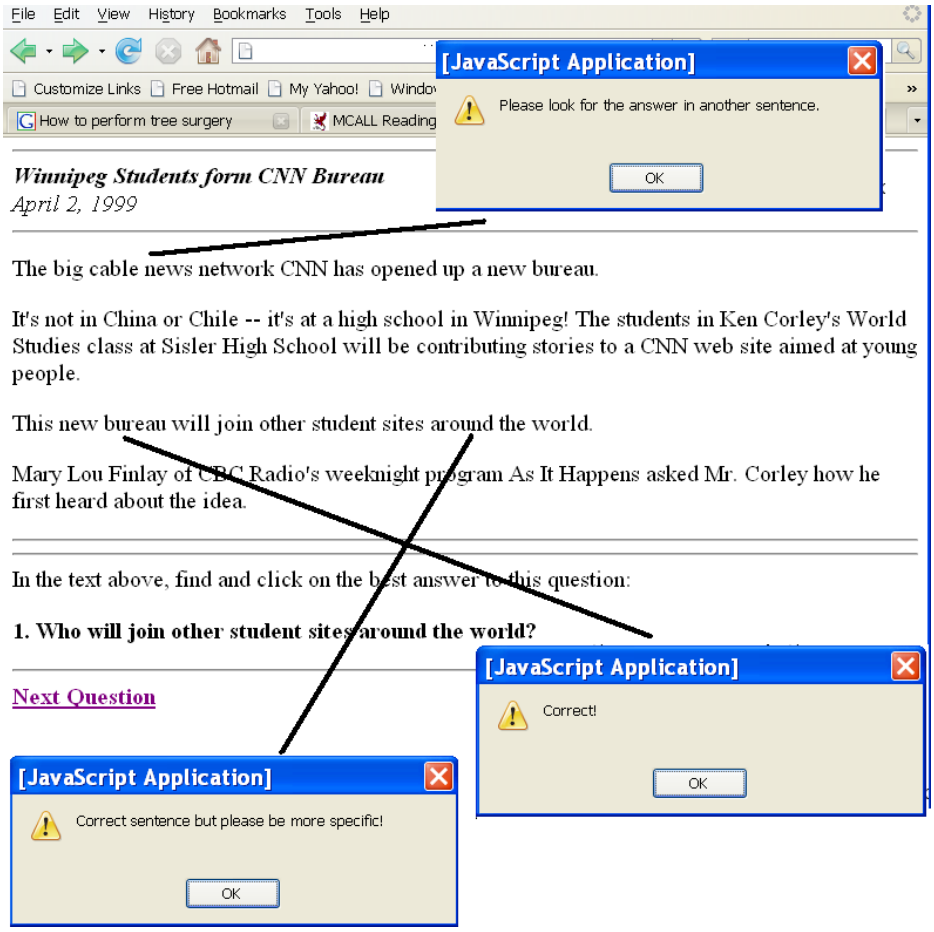
Wilson, E. (1997). The automatic generation of CALL exercises from general corpora. In Wichmann (Ed.), *Teaching and Language Corpora* (116-130). London; New York: Longman.

Wolfe, J. (1977). Automatic question generation from text - an aid to independent study. *Proceedings of the ACM SIGCSE-SIGCUE technical symposium on Computer science and education.*

Yoo, S., Gates, D., Levin, L., Fung, S., Arganwal, S., & Freed, M. (2008, June). *Using Semantic Features to Improve Task Identification from Email Messages*, NLDB 2008 (355-357). London, UK.

Appendices

Appendix A



A sample of various feedbacks given at different clickable points

in the reading passage for a specific WH question.