Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment

Gregory Aist

▶ To cite this version:

Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment

GREGORY AIST, Project LISTEN, 4215 Newell-Simon Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213 USA
E-mail: aist@cs.cmu.edu

Abstract. We address an important problem with a novel approach: helping children learn words during computer-assisted oral reading. We build on Project LISTEN's Reading Tutor, which is a computer program that adapts automatic speech recognition to listen to children read aloud, and helps them learn to read (http://www.cs.cmu.edu/~listen). In this paper, we focus on the problem of vocabulary acquisition. To learn a word from reading with the Reading Tutor, students must first encounter the word and then learn the meaning of the word from context. This paper describes how we modified the Reading Tutor to help students learn the meanings of new words by augmenting stories with WordNet-derived comparisons to other words – "factoids". Furthermore, we report results from an embedded experiment designed to evaluate the effectiveness of including factoids in stories that children read with the Reading Tutor. Factoids helped – not for all students and all words, but for third graders seeing rare words, and for single-sense rare words tested one or two days later. We also discuss further steps towards automatic construction of explanations of words.

INTRODUCTION

Project LISTEN's Reading Tutor listens to children read aloud, and helps them (http://www.cs.cmu.edu/~listen). The student and the Reading Tutor take turns choosing stories to read, including fiction or nonfiction selections. The Reading Tutor displays the story one sentence at a time, uses (adapted) speech recognition to listen to the child read all or part of the sentence aloud, and responds with help modelled in part after human experts. Children can use the Reading Tutor independently, in real classrooms (Figure 1).

The overall goal of the Reading Tutor is to help children learn to read. Ultimately, learning to read means learning to make meaning from print. Several bottlenecks present themselves along the way, for example sounding out new words, knowing the meaning of words, and integrating prior knowledge with new information from text. We focus here on one part of that overall goal: vocabulary acquisition – specifically, helping children learn the meaning of words. The core idea of this paper is: While a student is reading, explain unfamiliar words. Why? Children can learn vocabulary effectively from encountering words in everyday contexts, such as hearing stories read aloud (Robbins & Ehri 1994, Brett, Rothlein, & Hurley 1996), or reading text (Eller, Pappas, & Brown 1998; Nagy, Herman, & Anderson 1985). However, unannotated text alone may not provide enough information to learn much about the meaning of a word; adding information to text may help more than text alone. Rather than simply helping students understand text containing unfamiliar words – an assistive effect – we focus here on a learning effect: helping children learn new words. (Once you learn a new word, it's yours for a lifetime.)

Our long-term aim is to build an automated system that constructs a glossary for unrestricted text. Similar work includes systems for dynamically modifying link structure in hypertext, or generating hypertext (Brusilovsky et al. 1998). Sato (2001) describes a system for generating term explanations from the World Wide Web via information extraction, but the end product is several paragraphs long – partway between a dictionary entry and an encyclopedia article, and longer than some of the stories in the Reading Tutor. Here, we aim at augmenting an

existing document with dynamically generated (short) annotations, rather than generating the entire document from scratch, or generating comprehensive explanations of single words or phrases. In many cases, such as the ILEX system (Cox et al. 1999), there is a domain-independent engine that requires detailed data about a particular domain to generate concept explanations. We are taking a complementary approach: Discover and use domain-independent heuristics for when and how to give help, while keeping domain knowledge requirements minimal. Future systems could combine the knowledge-rich and knowledge-poor strategies, adapting their selection of help to fit the domain knowledge or other resources available.

Figure 1. A student reads with the Reading Tutor while the teacher teaches the rest of the class.

Reading material that contains new words is a requirement for learning new words from reading text. However, simply reading new and challenging stories may not be sufficient. Individual encounters with a word may not contain enough information to learn much about the word. We decided to explore augmenting text with vocabulary assistance. In the experiment described in this paper, we compared augmented text to unaugmented text, rather than to a "no exposure" control – because if the augmentation does not help over and above unaugmented text, adding augmentation would probably just waste the student's time.

We now turn to discussing several aspects of vocabulary help: which words to give help on, what kind of help to give, at what time to give help, and whether the computer or the student should decide whether to give help.

Which words should the Reading Tutor give help on? One option is to annotate all words. For example, the READER project (Schechter n.d.) annotated text by manually tagging each content word with its sense in WordNet (Fellbaum 1998), a computerized lexical database originally developed by George Miller and colleagues at Princeton. Another option is to specially prepare material. For example, one could hire people to write context-specific definitions for selected words. Another option is to design automated annotating, so that new text can also be augmented with assistance on words. We took this approach.

What kind of help should the Reading Tutor give? An example sentence is one option (Scott & Nagy 1997), but could break up the flow of reading by introducing extraneous ideas. For example, "The astronaut went to the Moon in a rocket" might cause confusion if inserted into a story that had nothing to do with going to the Moon. Another option is a conventional

definition. For example, "as·tro·naut. A person trained to pilot, navigate, or otherwise participate in the flight of a spacecraft" (American Heritage dictionary, 3rd edition, 1996). A better option for children is a definition from a children's dictionary. However, such definitions may vary widely in length and difficulty. For example: the definition for *astronaut* is short and sweet: "astronaut. a traveler in a spacecraft" (Merriam-Webster Student Dictionary, wordcentral.com). Consider, however, the definitions for *comet* and *meteor*: "comet. a bright heavenly body that develops a cloudy tail as it moves in an orbit around the sun"; "meteor. one of the small bodies of matter in the solar system observable when it falls into the earth's atmosphere where the heat of friction may cause it to glow brightly for a short time; also : the streak of light produced by the passage of a meteor" (Merriam-Webster Student Dictionary, wordcentral.com). A short explanation is another option: "An astronaut is someone who goes into outer space." Finally, a comparison to another word is even shorter and (hopefully) easier to read and understand: "An astronaut is a kind of traveler." We took this approach.

We wanted to add vocabulary assistance to text to make computer-assisted oral reading more effective for word learning. We did not intend to replace reading text with studying synonyms, as some previous studies have done (Gipe & Arnold 1978). Instead, we augmented assisted reading with comparisons to other words the student might already know. By analogy, consider salt: salt augments flavor, so salt is added to food – not used instead of food. Likewise, we did not contemplate completely replacing assisted reading with practice on synonyms – just augmenting text with semantic information to give students a learning boost when they encountered novel words. Furthermore, we did not intend to give a full definition of each target word – rather, supply some extra information aimed at supplementing the original text.

At what time should help be given? That is, when should the Reading Tutor provide vocabulary help on a word in a story – before the student reads the story, during the story, or after the story? For high school readers, Memory (1990) suggests that the time of instruction (before, during, or after the reading passage) for teaching technical vocabulary may not matter. If the lack of difference between presentation times holds true for elementary students as well, the Reading Tutor may be able to choose from several different times to give vocabulary help, without diminishing the student's ability to learn from the assistance. In the study described in this paper, we inserted vocabulary assistance just before the sentence containing the target word. (One study described in Aist (2000), Chapter 6, involved inserting vocabulary assistance before the story (limerick) containing the target word.)

Who should decide when help is necessary – the computer, the student, or both? We wanted to focus on the effects of vocabulary assistance unconfounded by whether the student requested help. Thus, in order to provide help equally to students who click frequently and to those who rarely click at all, we further chose to have the computer (or designer) control the presentation of words, rather than display explanations at the student's request.

In the remainder of this paper we describe an experiment on vocabulary assistance. We call this experiment the "factoid" experiment, because the type of vocabulary assistance we provided consisted of little facts about the target vocabulary words – or, factoids. The experiment had two main components. First, automatically generating and presenting factoids – that is, comparisons to other words (in the present study, drawn from WordNet). Second, automatically generating and administering assessments. We first give the five-step schema for the experiment, and then discuss each step in detail.

EXPERIMENT DESIGN

The design of the experiment described in this paper (Figure 2) was intended to contrast seeing a word in a story alone vs. seeing a word in a story along with some vocabulary help, as follows.

- **1. Student starts reading story,** with Reading Tutor assistance, up to just before the sentence containing the target word.

- **2. Reading Tutor (sometimes) provides a factoid.**
  2.a. If this is a control trial, nothing extra happens.
  2.b. If an experimental trial, the student reads a factoid (with Reading Tutor assistance).

- **3. Student continues reading story,** with Reading Tutor assistance.

- **4. Time passes.** One or more days pass.

- **5. Reading Tutor tests student's knowledge of the word** by presenting a multiple-choice vocabulary question at the start of the session on the next day the student logs in.
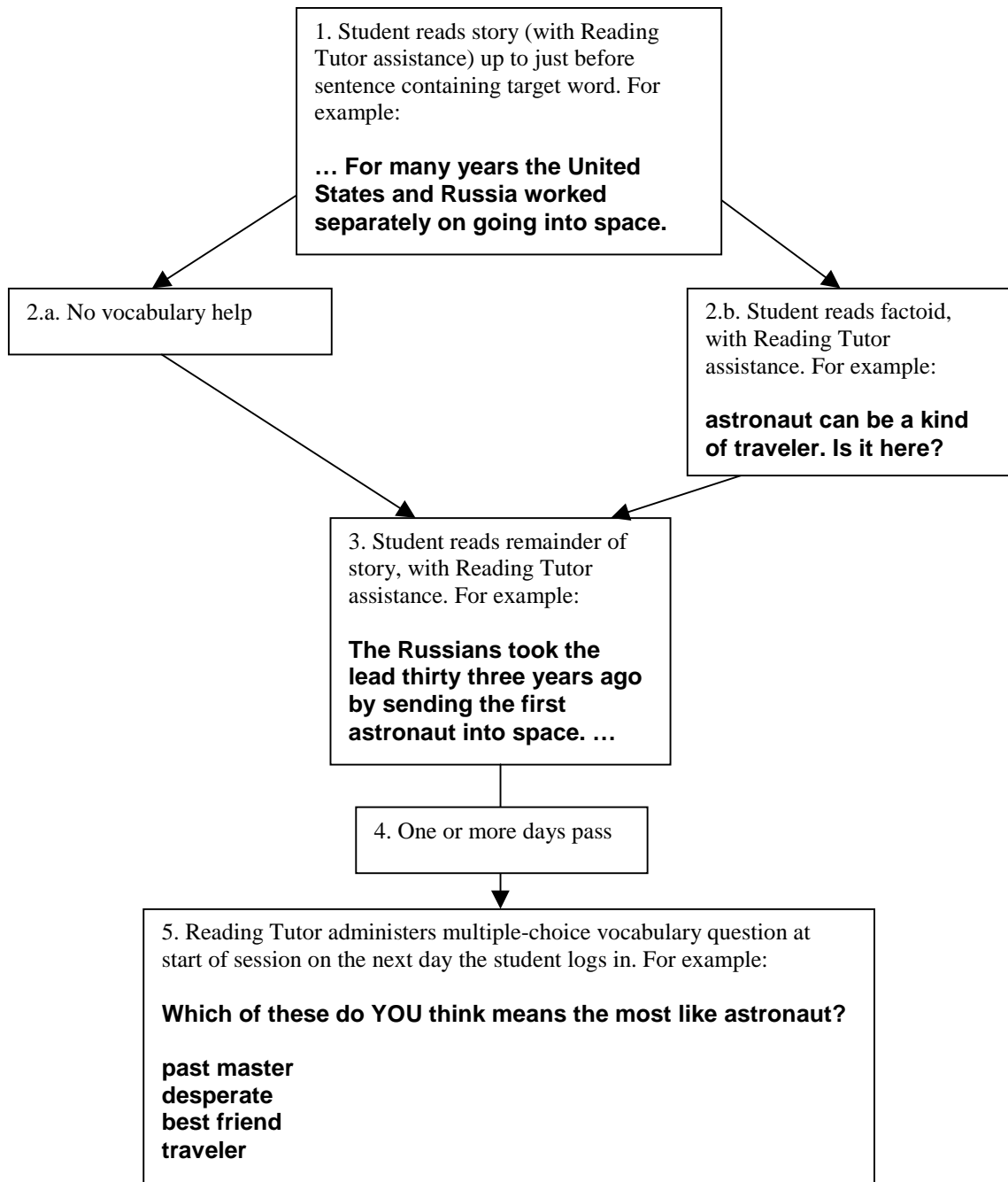

Figure 2. Factoid flowchart, showing one example using the target word *astronaut*. (*Desperate*, like the other three possible answers, is used as a noun here.)

Figure 2 shows the design of this experiment. We now discuss each step in turn.

Student starts reading story

A trial began with the student reading the story (with Reading Tutor assistance) up to just before the sentence containing the target word (Figure 3.) Basically, the Reading Tutor displayed one sentence at a time, listened to the student read aloud, and provided help on words it heard read incorrectly – words that the student may have missed or struggled with. The student could read a word aloud, read a sentence aloud, or read part of a sentence aloud.

The student could click on *Back* to move to the previous sentence, on the face or on the items in the *Help* balloon to request help on the sentence, or *Go* to move to the next sentence (Figure 3). The Reading Tutor moved on to the next sentence when it had heard the student read every content word (Aist 1997 provides details). The student could also click on *Goodbye* to log out.

The Reading Tutor responded when it heard mistakes or when the student clicked for help. Responses were constructed by playing hints or other help in recorded human voices. The help that the Reading Tutor provided sought to balance the student's immediate goal of reading the word or sentence with the longer-term goal of helping the student learn to read (Aist & Mostow 1997, Mostow & Aist 1999, Mostow & Aist 2001). Help included:

1. Read the entire sentence using a recording of a human narrator's fluent reading, to model correct reading. While playing the (continuous) recording, the Reading Tutor would highlight each word as it was spoken, which we call word-by-word highlighting.

2. Read the entire sentence by playing back isolated recordings of a single word at a time, in order to allow students to hear one word read at a time. Because these recordings may be in different voices, we call word-by-word playback "ransom note" help.

3. Recue a word by playing an excerpt from the sentence narration of the words leading up to that word (along with word-by-word highlighting), in order to prompt the student to try (re-) reading the word. For example: If the text is **Jack and Jill went up the hill to fetch a pail of water**, the Reading Tutor could recue **hill** by first reading **Jack and Jill went up the** out loud, and then underlining the word **hill** to prompt the student to read it.

4. Give a rhyming hint that matches both the sound (phoneme sequence) and the letters (grapheme sequence) of the target word, in order to give a hint on how to read the target word, and to expose the student to related words. For example, if the word is **hill**, give the word **fill** as a spoken and displayed rhyming hint, but not the word **nil** because its spelling does not match.

5. Decompose a word, syllable-by-syllable or phoneme-by-phoneme, to model the process of sounding out words and to call attention to letter-to-sound mappings. For example, say /h/ while highlighting **h**, then say /i/ while highlighting **i**, then say /l/ while highlighting **ll**.

6. Show a picture for a word, in order to demonstrate word meaning and to increase engagement. For example, if the word is **apple**, show a drawing of an apple. Fewer than 200 words had pictures in the 1997-1998 version.

7. Play a sound effect, perhaps to demonstrate word meaning but primarily to increase engagement. For example, if the word is **lion**, play the roar of a lion. Fewer than 50 words had sound effects in the 1997-98 version; most were names of animals, such as *seagulls*, *tiger*, and *wolf*.

Figure 3. Sentence before the sentence containing *astronaut*.

Reading Tutor (sometimes) provides a factoid

Providing factoids consisted of three steps: 1. selecting target words, 2. assigning target words to control (no factoid) or experimental (factoid) condition, and 3. generating and displaying a factoid. We describe each in turn.

Selecting target words.

During previous work, we had developed various constraints on the use of WordNet to provide help on unrestricted text. We devveloped these heuristics during the fall of 1998 by inspection of factoids generated to augment children's texts, such as *Alice of Wonderland*. Informal user testing over the summer of 1999 revealed further constraints, particularly the need to screen for social acceptibility (as described below.)

In 1999-2000, the Reading Tutor selected target words at runtime, using a set of predefined heuristics. A target word for factoid vocabulary assistance had to meet several conditions.

First, the Reading Tutor had to be able to give automated help on the word. We wanted to sidestep the challenge of word sense disambiguation on unrestricted text. Thus, the word had to have only a few senses in WordNet. Senses included those of the stemmed version (e.g. time) as well as from the actual text token (e.g. *times*). Stemming was done using WordNet's stemming function, called "morph". For a positive example: *astronaut* could be a target word because it has one sense: "astronaut, spaceman, cosmonaut -- (a person trained to travel in a spacecraft; 'the Russians called their astronauts cosmonauts')" (WordNet 1.6). For a negative example: *times* could not be a target word because while *times* has only the two senses "multiplication" and "best of times, worst of times", *time* has 14 senses.

Second, the Reading Tutor had to be able to ask a vocabulary question about the target word. We aimed at operationalizing Nagy et al.'s criterion of semantically similar distractors

(1985). To construct a 4-item multiple-choice vocabulary question, the Reading Tutor needed the correct answer and three wrong answers to serve as distractors – drawn from WordNet as described later.

Third, the word could not be a trivially easy word. The word must have been three or more letters long. The word could not be on a list of 36 words given by Mostow et al. (1994), shown in Table 1. In addition, the word must not have been a number written in Arabic numerals (for example, 200 or 35.)

Table 1. Thirty-six function words excluded from vocabulary experiment.

a	all	an	and	are	as	at	be	by	for	he	her
him	his	I	if	in	is	it	its	me	not	of	off
on	or	she	so	the	them	then	they	this	to	we	you

Fourth, the word could not be a proper noun. The word could not be a capitalized word (except for the first word in the sentence, which may be capitalized). This heuristic was designed to eliminate most names.

Fifth, the word had to have been socially acceptable. The target word, the comparison word, the intended answer, and the distractors all had to be socially acceptable. We screened for acceptability in two ways. To forestall obviously offensive words, we required a natural-speech narration of a target word to have been recorded beforehand by a Project LISTEN team member, since we trusted project members not to record inappropriate words. To exclude words that are fine to pronounce, but risky to give automatically generated semantic help on (such as words with secondary slang meanings), the word must not have been on a list of explicitly banned words. These heuristics avoided the most egregious problems – but they were not perfect, allowing phrases where each word in itself was inoffensive, but the entire phrase was not.

Assigning target words to conditions

We now describe how the Reading Tutor assigned words to conditions during the Fall 1999 factoid vocabulary study. For each student, half of the target words were randomly assigned to the experimental condition (factoid plus context), and the rest of the target words to a control condition (context alone). This randomization was done on a per-student basis. Thus while one student might see astronaut in the experimental condition, another student might see astronaut in the control condition. When the student encountered a previously unseen target word, the Reading Tutor assigned the new word to either the experimental (factoid plus context) or control (context alone) condition for that student. Since the same passages were used in control and experimental trials, this experiment controlled for text and word differences by randomly counterbalancing across trials, and relied on thousands of trials to wash out variance. While ultimately we might want to select words to explain based on what words are important to explain to which students, the Fall 1999 Reading Tutor used a blind, random assignment of words to conditions intended to persist for a given student's experience.

Providing a factoid

For a control word, the Reading Tutor did not provide a factoid – rather, it simply continued on to the next sentence. For an experimental word, the Reading Tutor displayed a factoid. How did it construct factoids?

We wanted to make vocabulary assistance that was applicable to any text. To do so, we needed a large-scale resource to cover many words students would encounter over the course of months of Reading Tutor use. We needed both to provide assistance and to assess its effects. To meet the goal of large-scale assistance and assessment applicable to any English text, we made use of a well-known lexical database: WordNet (Fellbaum 1998). WordNet contains tens of thousands of words organized by a thesaurus-style hierarchy (*astronaut* is a kind of *traveler*) and with links to synonyms (*astronaut* and *cosmonaut* are synonyms in WordNet). We designed

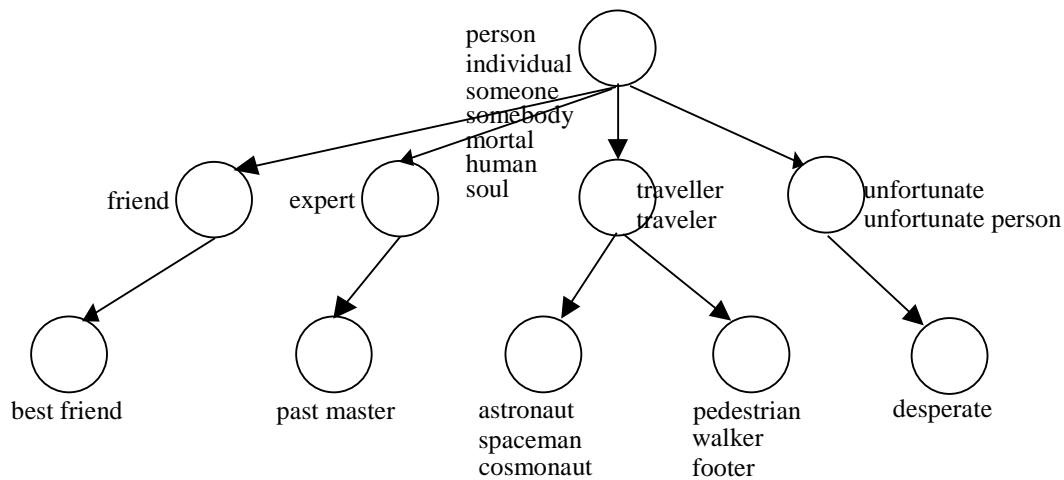automated assistance, applicable to any text, which compared words in the text to other words in WordNet.

Figure 4. Siblings and cousins in WordNet. Selected portion of the WordNet 1.6 hierarchy. Nodes contain the set of all words that are synonyms of each other – that is, that form a single synset. Arrows point from general to more specific nodes. astronaut is a sibling to pedestrian because their nodes share the parent node "traveller; traveler". astronaut is a cousin to best friend because their nodes share the grandparent node "person; individual; … soul".

The Reading Tutor displayed vocabulary help for target words in the form of short comparisons to other words. The other words were extracted from WordNet. The vocabulary help was hedged because it might have been incorrect. For example, the comparison word might have been related to a different sense of the word than actually appeared in the story. The hedge question also aimed to encourage the student to think about the meaning of the word in context. For example: "astronaut can be a kind of traveler. Is it here?"

The Reading Tutor constructed the text of the factoid from a template containing placeholders for the target word and for the comparison word. The templates used in the 1999-2000 study were as follows.

- Antonym. "The_Stem may be the opposite of The_Antonym. Is it here?"

- Hypernym. "The_Stem can be a kind of The_Hypernym. Is it here?"

- Synonym. "Maybe The_Stem is like The_Synonym here... Is it?"

Here, The_Stem was the base form of the word (*astronauts* → *astronaut*), The_Antonym was a word meaning the opposite of the target word, The_Hypernym was a more general word than the target word, and The_Synonym was a word that meant the same as the target word. Hypernyms and synonyms were used more frequently than antonyms. (Like many words, *astronaut* has no generally accepted opposite.) Figure 4 shows an excerpt of the WordNet hierarchy containing *astronaut*.

To make sure that the student would pay attention to the vocabulary assistance, and to give the student extra practice in reading the target word, we presented the vocabulary assistance as text for the student to read out loud with the Reading Tutor's help. (Other possibilities we considered included simply speaking the vocabulary assistance, presenting the text briefly in a drop-down window below the original sentence, or some combination of spoken and drop-down text.)

We also had a number of other design goals which were met in extended joint design work with the present author and Project LISTEN team members, especially Kerry Ishikazi and Jack Mostow; also Human-Computer Interaction Masters' student Margaret McCormack.

- To distinguish the factoid from the original text, we placed the factoid on a yellow background.

- To attribute the factoid to the Reading Tutor instead of the author of the original text, we placed the factoid in a call-out balloon attached to the face in the lower left hand corner of the screen.

- To avoid confusion about what to read, and to simplify layout, the balloon occludes the original text.

- To provide first-class assistance, the factoid is presented as text for the student to read aloud, with Reading Tutor assistance (Presenting the factoid as text to read also allowed for the possibility of giving factoids on factoids – we didn't, but might want to in the future.)

Figure 5 shows the factoid as displayed by the Reading Tutor.

Figure 5. Factoid in popup window.
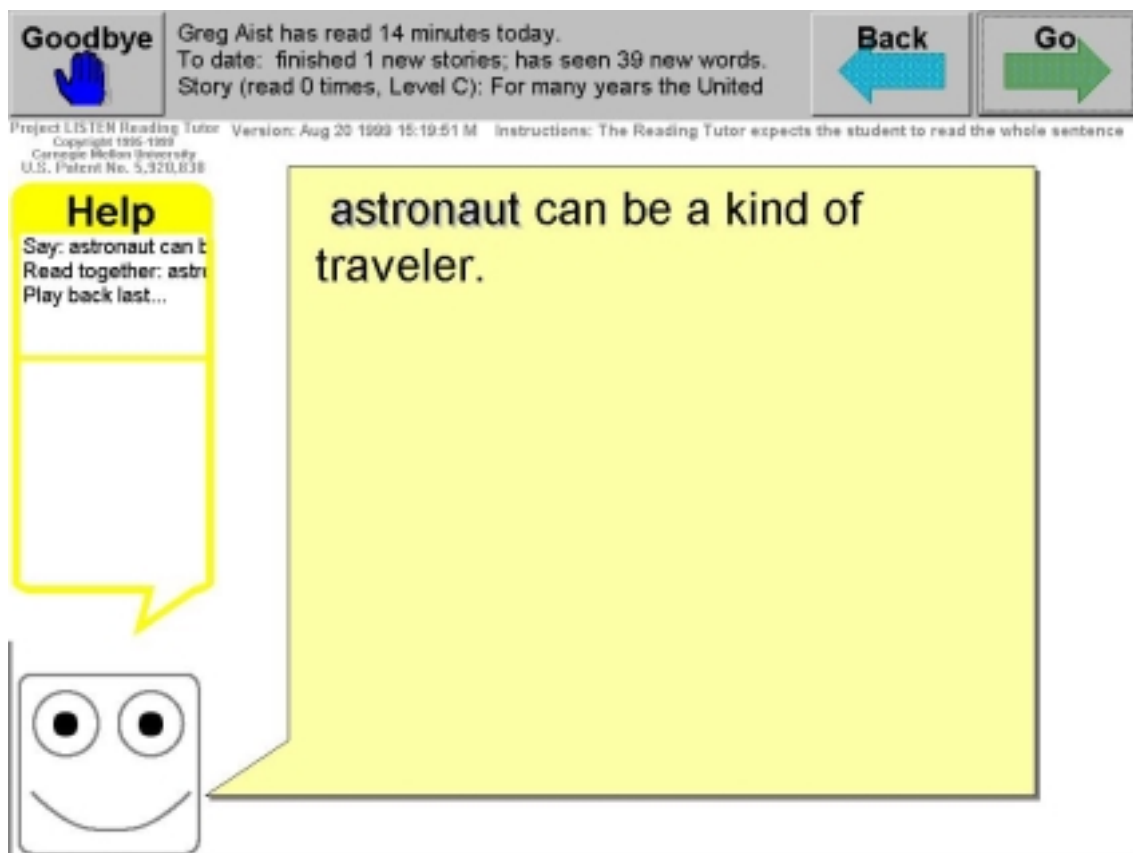
Student continues reading the story

After reading the factoid (or not reading it, for control words), the student continued to read the story with the Reading Tutor's assistance. Figure 6 shows the sentence containing *astronaut*.

Time passes

One or more days went by. On a subsequent day, the student logged in again as usual to begin working with the Reading Tutor.

Reading Tutor tests student's knowledge of the word

In order to test the effects of this assistance, the Reading Tutor administered multiple choice questions on a later day. We now describe in detail how the Reading Tutor generated vocabulary assistance.

Figure 6. Sentence containing the word *astronaut*.

We needed to evaluate the effectiveness of vocabulary assistance. Nagy, Herman, and Anderson (1985) categorized multiple-choice questions according to how close the distractors (incorrect answers) are to the correct answer. Nagy, Herman, and Anderson's classification is as follows:

- Level 1. Distractors are a different part of speech from the correct answer. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 1 distractors might be *eating*, *ancient*, and *happily*.

- Level 2. Distractors are the same part of speech but semantically quite different. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 2 distractors might be *antelope*, *mansion*, and *certainty*.

- Level 3. Distractors are semantically similar to the correct answer. For example, if the target word is astronaut and the correct answer is *traveler*, Level 3 distractors might be *doctor*, *lawyer*, and *president*.

We designed automated vocabulary assessment questions using the WordNet hierarchy, taking as our goal Nagy, Herman, and Anderson's Level 3 multiple choice questions.

The Reading Tutor used synonyms and hypernyms as the correct answer, reverting to a sibling (Figure 4) only if neither a synonym nor a hypernym could be found.

Vocabulary question distractors were cousins of the target word (words with a common grandparent but different parents) (Figure 4.) The distractors were chosen so that the multiple

choice question tested a student's ability to select the meaning of the target word from several semantically similar alternatives.

Skeptics might ask: Why use automated, variable-quality, experimenter-defined questions instead of a standardized instrument? First, our decision to use experimenter-defined questions was subsequently validated by the National Reading Panel's (later) call for the use of experiment-defined measures to test vocabulary. Experimenter-defined measures, which often measure gains on particular words, tend to be more sensitive to small gains than standardized tests that aim to measure vocabulary in general by sampling a small number of words (NRP 2000). Our questions were experimenter-defined in the sense that we wrote code which generated the questions – as opposed to using an off-the-shelf vocabulary assessment. Second, we used this measure as a comparison inside the interaction. That is, all students received assistance, saw words in the experimental and control conditions, and took the multiple-choice tests. Thus our experiments were within-subject in a way that standardized tests do not (easily) facilitate. Third, we did in fact pre-test the students using a widely used measure of vocabulary – the Word Comprehension subtest of the Woodcock Reading Mastery Test (WRMT, see Aist 2000 Chapter 5 for details). As a test of external validity, we calculated the correlation between students' performance on the multiple-choice questions for words seen in context (without extra assistance), and their (grade-normed) performance on the Word Comprehension section of the WRMT. The correlation was significant, at $r=0.47$ for grade 2 ($p = .009$) and $r=0.49$ for grade 3 ($p = .008$). Thus our measure fit the national research agenda, could be automatically constructed and scored, and was correlated with a widely used vocabulary assessment.
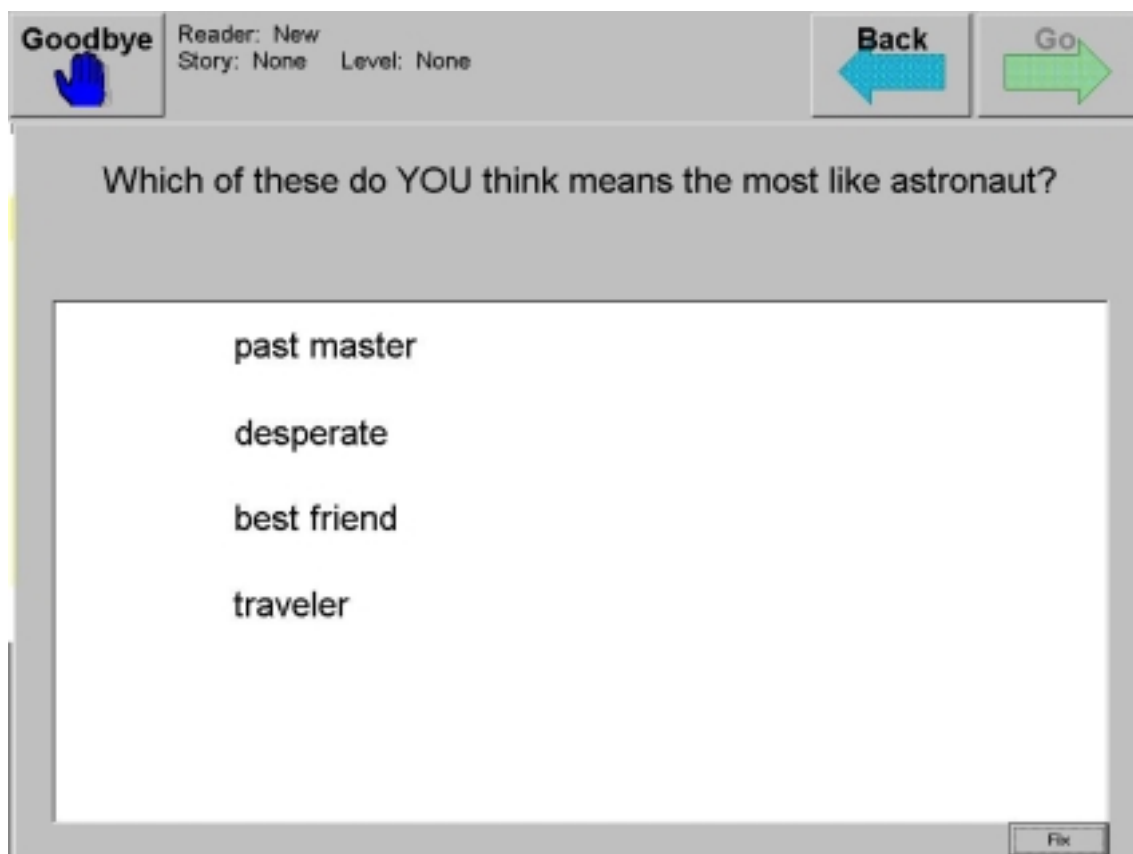
Figure 7. Multiple-choice question for factoid experiment. The informally phrased prompt was written to make sense for the varied kinds of target words: *astronaut*, *twinkling*, *Rwanda*, etc.

We assessed the effectiveness of vocabulary intervention as follows. The next time a student logged in (one day or more after seeing the target word) the Reading Tutor displayed a vocabulary question for each of the target words the student had encountered – both experimental and control words. The answers were displayed in a random order. (Random order

was a potential source of variance, but reduced possible effects of children seeing their peers answering questions on the same words, when it was their turn to use the Reading Tutor.) Also, the selection of a particular correct answer and distractors was not constant for a word, but chosen anew for each trial. The Reading Tutor spoke the prompt at the top of the screen, and then spoke the answers one at a time while highlighting each answer in yellow. The student could select an answer at any time by clicking on it; nonetheless the vocabulary questions did take time to answer (ranging from 14-21 seconds each in the example given in the next section). Since the vocabulary questions were administered at the start of the session, they could not be contaminated by additional assistance just before being asked.

AN EXAMPLE OF AN EXPERIMENTAL TRIAL

Here is an example of an experimental trial, excerpted from actual Reading Tutor use during Fall 1999. We display events involving the target word *astronaut* in boldface.

	Time event occurred	What happened?
	Wednesday, October 6, 1999 12:37:10.356	Student (P.O., girl aged 9 years 5 months) chooses Level C story "Life in Space" (adapted from a Weekly Reader passage)
2 seconds later	12:37:12.259	Reading Tutor displays sentence "For many years the United States and Russia worked separately on going into space." Student tries reading sentence out loud.
19 seconds later	12:37:31.106	Student finishes speaking. Actual utterance: for many years the united states of russia worked s... sponidy on going to space Reading Tutor heard: FOR MANY YEARS THE UNITED STATES AND RUSSIA WORKED SEPARATELY SEPARATELY ON GOING INTO SPACE (The Reading Tutor's hearing is not perfect; in this case, it may have not detected the miscue because "sponidy" sounded more like "separately" than like the other words in the sentence (or truncations thereof), which is all the Reading Tutor listened for.)
< 1 second later	12:37:31.166	Reading Tutor decides to display next sentence of story
24 seconds later	**12:37:55.391**	**Reading Tutor displays first sentence of factoid: "astronaut can be a kind of traveler." Student tries reading sentence.**
16 seconds later	12:38:11.464	Student finishes speaking; Reading Tutor heard: ASTRONAUT CAN BE A KIND OF TRAVELER ASTRONAUT CAN BE A KIND OF TRAVELER
< 1 second later	12:38:11.524	Reading Tutor decides to go on to the next sentence
3 seconds later	12:38:14.408	Reading Tutor displays second sentence of factoid: "Is it here?"
9 seconds later	12:38:23.571	Student finishes speaking; Reading Tutor heard: IT INDIA IS IT HERE (What the Reading Tutor heard was not necessarily what the student actually said. If the sentence was short, the Reading Tutor included additional words to listen for, to approximate students' oral reading insertions and deletions, and to reduce acceptance of incorrect student attempts. Here, one "extra" word was INDIA.)
< 1 second later	12:38:23.621	Reading Tutor decides to display next sentence
1 second later	**12:38:24.843**	**Reading Tutor displays: "The Russians took the lead thirty three years ago by sending the first astronaut into space."**
		[Time passes]
Almost 24 hours later	Thursday, October 7, 1999 12:28:06.621	Student logs in the next day
2 seconds later	12:28:08.564	Reading Tutor presents student's name, for student to read as confirmation of identity and to make sure the microphone was working
10 seconds later	12:28:18.098	Student finishes reading name

9 seconds later	12:28:27.581	Reading Tutor presents vocabulary question by displaying the question and the answers, reading the question and then the answer out loud. Which of these do YOU think means the most like pail? railway car; paper bag; bucket; piles
16 seconds later	12:28:43.845	Student clicks on *bucket* (right!)
6 seconds later	12:28:49.713	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like asparagus? butterfly pea; bog plant; yam plant; herb
20 seconds later	12:29:10.232	Student clicks on *herb* (right!)
17 seconds later	**12:29:36.881**	**Reading Tutor presents vocabulary question:** **Which of these do YOU think means the most like astronaut?** **past master; desperate; best friend; traveler**
17 seconds later	**12:29:54.025**	**Student clicks on *traveler* (right!)**
5 seconds later	12:29:59.013	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like fetch? bring; project; impact; ferry
14 seconds later	12:30:13.073	Student clicks on *impact* (wrong!)
4 seconds later	12:30:17.299	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like silk? material; hill; piece; cross
21 seconds later	12:30:37.708	Student clicks on *material* (right!)
8 seconds later	12:30:45.760	Reading Tutor chooses Level A story: "The Letter A"

A few notes on this example: First, displaying factoids sometimes caused delay due to database access. (In the case of astronaut in this example, 24 seconds). Second, it was not unusual for students to repeat a sentence if the Reading Tutor did not immediately accept their reading ("ASTRONAUT CAN BE A KIND OF TRAVELER ASTRONAUT CAN BE A KIND OF TRAVELER"). Finally, we discuss later in the paper such oddities in the factoids and questions as *asparagus* being a kind of *herb*.

The Reading Tutor showed one factoid for every experimental trial. Thus, if two target words were in a single sentence, and both target words were randomly assigned to the experimental condition, the Reading Tutor would show a separate factoid for each target word.

The data used in this paper were collected from second and third graders' use of the Reading Tutor in Fall 1999 at an elementary school near Pittsburgh, Pennsylvania. (Students actually used the Reading Tutor during the whole school year, but this experiment was only active during Fall 1999.)

A minor bug caused the Reading Tutor to display multiple factoids for certain words, namely just those words which occurred as the first word of the sentence, capitalized. These few words (less than ten) were excluded from the analysis of the experiments.

By having an open-ended set of target words instead of a fixed list, we enabled the Reading Tutor to give assistance without disrupting the study design on any new material added by teachers, students, or the Project LISTEN team. The assignments of words to conditions were intended to persist throughout a particular student's history of Reading Tutor use to enable us to look for longer-term effects of multiple exposures to a word. Unfortunately, due to a flaw in the software, the assignments were not saved to disk. We therefore analyzed only a student's first day of experience with a word, and the subsequent vocabulary question.

We used a database to collect the data from the over 3000 factoid trials. One trial was not properly recorded to the database due to a full hard drive: on Wednesday, March 22, 1999, a student received help on the word *pounce* that was recorded in the Reading Tutor's log file, but not in its database.

RESULTS AND ANALYSIS

How much did factoids help? In order to assess the effectiveness of factoid assistance overall (3359 trials), we compared student performance on the experimental condition (factoid + context, 1679 trials) to student performance on the control condition (context alone, 1680 trials).

Individual students' performance on all conditions ranged from 23% to 80%, with chance performance at 25% (1 out of 4).

The (U.S.) National Reading Panel Report, a consensus expert survey of research on how to help children learn to read, remarked that many reading studies choose the wrong value of N when conducting analyses (NRP 2000). In this case, analyzing the factoid experiment using the trials as independent data points would be statistically incorrect, because a given student's trials were not independent of one another, and also because the number of trials varied from student to student. Analyzing the factoid experiment by direct comparison of per-student averages would underestimate the effective sample size, because the average is not a single measure but rather a composite of multiple related trials.

Logistic regression models offer a statistical technique for representing multiple responses from multiple students, and analyzing the results. Thus, to explore the effect of factoids on getting the question right, we built logistic regression models using the statistical software package SPSS. Logistic regression predicts a binary outcome variable using several categorical factors, and is a statistically valid technique for analyzing experiments with multiple responses per student – more sensitive than analysis of variance over the mean of students' answers, and more statistically appropriate than paired T-tests over all answers. See Menard (1995) for further information about logistic regression. Here, the outcome variable was whether the student got the answer right or not. The following factors were included in the model:

- whether the student received a factoid on the target word;

- who the student was, so as to prevent bias towards students with more trials, and to properly group a student's trials together;

- a term for how difficult the questions were overall – that is, background difficulty; and

- a term for what the effect of help was on getting the question right.

If the coefficient for the effect of help on getting the question right was (significantly) greater than zero, then factoids (significantly) boosted student performance. We accompany the description of results below with figures on average percent correct, calculated on a per-student basis to avoid bias towards students who encountered more target words.

No significant effect overall

Did factoids significantly boost performance? The per-student average percentage correct for the control trials was 37.2% with standard deviation 16.9%; for experimental trials, 38.5% with standard deviation 18.3%. (Per-student percentages have high standard deviations because they are averages of individual rates, which vary by student.) The coefficient for the effect of help on getting the question right was 0.07 ± 0.07, for all 3359 trials. Thus, factoids did not significantly boost performance overall – due perhaps to a number of problems with automated assistance that we next sought to filter out.

Exploration revealed possible effect for rare, single-sense words, tested 1-2 days later

The overall analysis looked at the effect of (imperfect) factoids as measured by (imperfect) multiple-choice questions, "warts and all" if you will. To get at the question of what would have happened without (some of) the warts, we decided to examine conditions under which the factoids might have been effective. The exploratory nature of the following analysis means that its results should be considered suggestive, not conclusive. What conditions might affect the effect of factoids?

Some words in the target set had more than one meaning. Students might well be confused – or at least not helped – by factoids that explained a different sense of the target word than was used in the original text. Perhaps factoids were effective only for single-sense words. Did factoids help for single-sense words only? Not significantly, but the trend was still positive (Table 2).

Table 2. Single-sense difficult words in the factoid experiment.

Total number of trials, from how many students	How were trials selected?	Per-student average number right	Coefficient in logistic regression model
720 trials 59 students	Single-sense words	34.9% ± 23.0% for control vs. 38.4% ± 26.5% for experimental	0.23 ± 0.17
191 trials 52 students	Single sense words coded as hard by a certified elementary teacher	26.3% ± 30.0% for control vs. 29.1% ± 36.8% for experimental	.13 ± .41
348 trials 55 students	Single sense words coded as hard by the experimenter	33.0% ± 29.0% for control vs. 40.7% ± 36.3% for experimental	.35 ± .27
317 trials 55 students	Single sense rare words	35.4% ± 30.5% for control vs. 42.4% ± 37.3% for experimental	.16 ± .29
189 trials 48 students	Single-sense rare words tested one or two days later	25.8% ±29.4% for control vs. 44.1% ± 37.7% for experimental	1.04 ± .42 Significant at 95%, exploratory and thus not correcting for multiple comparisons

Some of the words in the target set were easy – *apple*, for example. Presumably, if a student already knew a word, a factoid would not help. Did factoids help for single-sense hard words? Maybe. We manually classified each target word as hard or not hard. So as to avoid biasing the classification due to knowledge of the outcome of the trials, we classified the words without looking at the outcomes on individual trials or words. We also identified the words that were rare – words that occurred fewer than 20 times in the million-word Brown corpus (Kucera and Francis 1967). Results were again not significant, but suggestive of a positive impact of factoids, as follows (Table 2). (The fact that the highest performance was only 44.1% suggests that there is room for improvement in the factoids, the questions, or most likely in both.)

Perhaps students learned or remembered enough of the help to do better a few days later, but not over an extended period of time such as a weekend. Did the factoids help for single-sense rare words tested one or two days later? Yes (Table 2). If the effect only persists for a few days, how could we improve students' retention of the meanings they learned? Future work might aim at reinforcing this retention with a second exposure to the target word.

As a sanity check, we looked at the 27 words in these trials (Table 3). Most of the words in Table 3 seem plausible as words that some elementary school students might not know, and for which explanations might be helpful. Selecting trials where the test occurred only one or two days after the training meant including fewer trials from students who were frequently absent, introducing a self-selection bias. Therefore, we next explored the factoid results using attributes that did not reflect self-selection, but rather other properties of the students such as grade.

Table 3. Single-sense rare words tested one or two days later.

Word	Word frequency in Brown corpus	Example of a factoid	Example of a multiple-choice question – correct answer in **bold**, student's response underlined.
aluminum	18	aluminum can be a kind of metal	**Al**;wood coal;soot;<u>black lead</u>
astronaut	2	astronaut can be a kind of traveler	<u>married person</u>; **traveler**; computer; nerd
bliss	4	Maybe bliss is like walking on air here	seeing red; scare; **walking on air**; <u>melancholy</u>
bobbin	-	Maybe bobbin is like reel here	**<u>reel</u>**; wheel; power train; gun
coward	8	coward can be a kind of mortal	<u>Old Nick</u>; young; **someone**; escape
crouching	-	crouch can be a kind of sit down	**<u>crouched</u>**; lace; wring; mat
daisies	-	daisy can be a kind of flower	prairie star; painted cup; snow plant; **<u>flower</u>**
eggshell	1	eggshell can be a kind of natural covering	Little Dog; mouth; meat; **<u>cover</u>**

glittering	-	-	stay together; **shine**; carry; lurch
headdress	-	headdress can be a kind of apparel	plastic wrap; **wearing apparel**; <u>dust cover</u>; arm
hello	10	Maybe Hello is like hi here	<u>good day</u>; good night; morning; **hi**
infirmities	-	infirmity can be a kind of bad condition	sore; **bad condition**; <u>wound</u>; twist
liar	3	liar can be a kind of cheat	<u>runner</u>; **cheat**; beast; wolf
outskirts	-	outskirt can be a kind of city district	hub; nation; **city district**; <u>roads</u>
pasta	-	pasta can be a kind of food product	**food product**; chow; <u>bird food</u>; food cache
pebbles	-	pebble can be a kind of stone	clay; **rock**; sheath; Crow
plat[1]	-	plat can be a kind of map	**map**; chalk; check; rule
plumage	-	plumage can be a kind of animal material	<u>mineral pitch</u>; **body covering**; dye; winter's bark
pollen	11	Pollen can be a kind of powder	**<u>powder</u>**; diamond dust; water glass; milk glass
princess	10	Princess can be a kind of blue blood	coach; chair; **blue blood**; mayor
Rwanda	-	Rwanda can be a kind of African country	**<u>African country</u>**; England; United Kingdom; United States
salad	9	salad can be a kind of dish	bite; **dish**; <u>breakfast</u>; choice morsel
tennis	15	tennis can be a kind of court game	field game; <u>night game</u>; **court game**; day game
twinkling	2	-	ping; **second**; ring; <u>bang</u>
vales	-	Maybe vale is like valley here	Blue Ridge Mountains; hill; **valley**; <u>bank</u>
wading	-	wade can be a kind of walk	work; pace; bounce; **<u>walk</u>**
wayside	2	wayside can be a kind of edge	arm band; **margin**; strap; <u>ring</u>

Further characterization of factoid results

In order to more fully characterize the factoid results, we looked at a number of possible subdivisions of the data with respect to their effects both on the percentage of correct answers, and on the coefficient for effect of factoid on answer in the regression model. Table 4 shows percentage correct – calculated as the average of the per-student mean – and the effect of factoid on answer for several subdivisions of the data.

Table 4. Further characterisation of the factoid results.

Which students?	Which words?	Trials	Percentage correct	Outcome: Coefficient ± 1 s.d.
All students	All words	3359	37.2% ± 16.9% control 38.5% ± 18.3% expt.	No effect of factoid: 0.07 ±- 0.07
33 students in Grade 2	All words	1391	35.4% ± 11.7% control 33.1% ± 11.6% expt.	No effect of factoid: -0.03 ±- 0.12
36 students in Grade 3	All words	1968	33.1% ± 11.0% control 42.0% ± 19.0% expt.	Trend favoring factoid: 0.15 ± 0.10
All students	Single-sense	769	36.8% ± 26.6% control 39.2% ± 29.2% expt.	Slight trend favoring factoid: 0.21 ± 0.17
All students	Multiple-sense	2605	37.4% ± 17.2% control 37.3% ± 20.2% expt.	No effect of factoid: 0.07 ± 0.08
All students	Rare words	1927	35.6% ± 19.5% control 38.3% ± 21.1% expt.	No effect of factoid: 0.13 ± 0.10
All students	Non-rare words	1427	40.0% ± 18.6% control	No effect of factoid:

[1] *plat* appeared in an (apparently) student-written story which included the sentence "slapt flash slise plair clio ciay glass plat".

			37.8% ± 22.3% expt.	-0.06 ± 0.11
Grade 3	Rare words	465	36.2% ± 22.9% control 42.0% ± 28.4% expt.	Effect of factoid: 0.37 ± 0.21, p < .10
Students below median on weighted score of WRMT word comprehension pretest	All words	1319	33.1% ± 11.1% control 32.9% ± 9.4% expt.	No effect of factoid: 0.07 ± 0.12
Students at or above median on weighted score of WRMT word comprehension pretest	All words	1852	38.3% ± 9.7% control 42.3% ± 16.6% expt.	No effect of factoid: 0.11 ± 0.10

Word recency effect

The comparison word (traveler in "astronaut can be a kind of traveler") and the expected correct answer were drawn from partially overlapping sets of words. Because of the overlap between sets, 993 out of the 1709 experimental trials in this experiment used the same word for the comparison word and the expected answer, and the other 716 used a different word. The effects found when analyzing all of the trials could be due solely to a recency effect from having seen the comparison word on a previous day. Later experiments on augmenting text with definitions were designed to avoid such recency effects (Aist 2001).

DISCUSSION AND FUTURE WORK

There are two main avenues of future work towards automatic glossarization. First, in the present study, students most likely already knew some of the words. This problem could be addressed by student modeling: not only tracking words students have encountered previously, but using word frequency and perhaps other factors to identify words that a particular student could benefit from having explained.

The second problem is to improve the quality of the factoids: some of the factoids were not helpful since they were too hard, too obscure, or didn't match the sense of the word used in the original text. There were also various problems with the automated assessment that may have obscured the effectiveness of factoids. Some comparison words may have been harder to understand than the target words. Some of the incorrect answers (distractors) in the vocabulary questions were themselves rare – such as *butterfly pea* – making the question hard to understand. Or, questions may have relied on uncommon knowledge, such as *asparagus* being (botanically) an herb. In addition, hypernyms did not always capture the key elements of the meaning of a word. An *expert* may be a kind of *mortal*, but that fact, while true, is not terribly useful for a beginning reader. Even a reasonable substitution like *traveler* for *astronaut* leaves out the key concept (and word) *space*. Similarly, saying that *wade* is a way to *walk* leaves out the key concept (and word) *water*. Instead, we might want to define a new lexical relation, *explains*, where one word *foo* explains another word *bar* if (a) *foo* is an easier word than *bar* and (b) *foo* means the same as *bar*. Specifying exactly what *means* means here is itself difficult; one could use a conventional synonym relationship, or perhaps rely on word associations elicited during psychological studies such as that in the Edinburgh Associative Thesaurus (Kiss et al. 1973), or perhaps rely on corpus or Internet data to pair up words which occur in similar contexts. While *explains* would not cover all words, extension to short phrases might cover many words, for example: a *coward* is a *fearful person*; to *wade* is to *walk through water*.

Another question is, why automatically generate factoids at all? An 80/20 solution where factoids (or explanations) were generated automatically and then hand-filtered offers a reasonably cost-effective alternative to fully automated methods. However, such fixed resources would lack the (potential) adaptability of automatically generated explanations, which could for example take into account what words a student has seen before in order to relate a new word to not just easier words, but words that a particular student has previously learned. However, as a reviewer pointed out, effective factoids apparently must be better constructed, and these more intelligent factoids will require further work: selecting words with an eye to not just difficult words but towards those words which are either important to understanding the passage at hand, or in addition for use in later reading; replacing or augmenting WordNet with some other source of comparison words; better natural language generation to replace (or supplement) templates.

Questions of initiative and modality are natural avenues for follow-on studies. For the purposes of this study, we had the computer decide when and how to present factoids. In a non-experimental system, a mixed-initiative approach makes sense: for some words, the computer could volunteer explanations, while for others, the computer would not volunteer them but the student could request them. In addition, for this study we always used narrated text as the presentation modality. Future studies might explore alternative modalities for different students – perhaps audio recordings for younger students, or diagrams for older students. Finally, future experiments could employ (improved) automated help, but use manually written multiple-choice questions (or some other question format, perhaps free-input) for evaluation, to avoid the problems we encountered with (noisy) automatically generated questions.

CONCLUSION: LESSONS LEARNED FROM FACTOID STUDY

At the end of Fall 1999, we turned off the vocabulary questions for a number of reasons. The primary reason was because they were getting slower and slower as database queries labored over data collected during the entire year to date, a problem which we feared was excessively frustrating. Our decision was however also due to the problems with factoids and questions we have already discussed. We did however leave the factoids on, to avoid excessive changes to what children did on the Reading Tutor. Turning vocabulary questions off precluded carrying out fine-grained analysis of factoids in Spring 2000. However, elsewhere we present a more summative analysis: results pertinent to vocabulary learning from a year-long evaluation of the Reading Tutor with Take Turns and factoids, which compared the Reading Tutor to classroom instruction, and also to one-on-one human tutoring (Mostow et al. 2001, Aist et al. 2001) The short version: Computer-assisted oral reading helps third graders learn vocabulary better than a classroom control – about as well as human-assisted oral reading. The long version (along with other results) was under preparation as a separate manuscript at press time.

Factoids helped – sometimes – but generating good assistance automatically requires common sense that code lacks. Thus, at least for the near term we recommend using vocabulary assistance as follows: either constructed by machine and then hand filtered, or directly constructed by hand. (We followed our own advice, as the designs of subsequent experiments bear out (Aist 2000), where we augmented children's limericks with hand-written definitions of difficult words.). Nonetheless, the factoid study suggests that augmenting text with factoids can help students learn words, at least for third graders seeing rare words ($p < .10$), and for single-sense rare words tested 1-2 days later ($p < .05$).

Acknowledgements

As with all research carried out within the context of a larger project, the present paper was enabled by previous work done by many on Project LISTEN; the project website lists personnel (http://www.cs.cmu.edu/~listen). This paper is an extended version of Aist (2001); we thank anonymous AI-ED and IJAIED reviewers for their comments, Brian Junker for statistical advice, and Jack Mostow and Brian Tobin for reading and commenting on earlier drafts of this paper. Any remaining problems are of course the sole responsibility of the author.

References

Aist, G. (1997). Challenges for a mixed initiative spoken dialog system for oral reading tutoring. AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interaction. AAAI Technical Report SS-97-04.

Aist, G. (2000). Helping Children Learn Vocabulary During Computer-Assisted Oral Reading. Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, December 2000.

Aist, G. (2001). Factoids: Automatically constructing and administering vocabulary assistance and assessment. *Proceedings of the 10th International Conference on Artificial Intelligence in Education (AI-ED 2001)*. San Antonio, Texas, May 19-23, 2001.

Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Junker, B., Mostow, J., Sklar, M.B., and Tobin, B. (2001). Computer-assisted oral reading helps third graders learn vocabulary better than a classroom control – about as well as human-assisted oral reading. *Proceedings of the 10th International Conference on Artificial Intelligence in Education (AI-ED 2001)*. San Antonio, Texas, May 19-23, 2001.

Aist, G., and Mostow, J. (1997). Adapting human tutorial interventions for a reading tutor that listens: using continuous speech recognition in interactive educational multimedia. In *Proceedings of CALL 97: Theory and Practice of Multimedia in Computer Assisted Language Learning*. Exeter, UK.

Brett, A., Rothlein, L., and Hurley, M. (1996). Vocabulary acquisition from listening to stories and explanations of target words. *The Elementary School Journal* 96(4), 415-22.

Brusilovsky, P., Kobsa, A., and Vassileva, J. (eds.) (1998.) *Adaptive Hypertext and Hypermedia*. Dordrecht: Kluwer Academic Publishers.

Cox, R., O'Donnell, M., and Oberlander, J. (1999). Dynamic versus static hypermedia in museum education: an evaluation of ILEX, the intelligent labelling explorer. *Proceedings of the Conference on Artificial Intelligence in Education* (AI-ED 1999), Le Mans, July 1999.

Eller, R. G., Pappas, C. C., and Brown, E. (1998). The lexical development of kindergarteners: Learning from written context. *Journal of Reading Behavior* 20(1), 5-24.

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

Gipe, J. P., and Arnold, R. D. (1978). Teaching Vocabulary through Familiar Associations and Contexts. *Journal of Reading Behavior* 11(3), 281-285.

Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), *The Computer and Literary Studies*. Edinburgh: University Press.

Kucera, H., and Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Memory, D. M. (1990). Teaching technical vocabulary: Before, during or after the reading assignment? *Journal of Reading Behavior* 22(1), 39-53.

Menard, Scott. (1995). *Applied Logistic Regression Analysis*. vol. 106, Quantitative Applications in the Social Sciences. Sage Publications.

Mostow, J., and Aist, G. (1999). Giving help and praise in a Reading Tutor with imperfect listening – Because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3), 407-424. Special issue (M. Holland, Ed.), Tutors that Listen: Speech recognition for Language Learning.

Mostow, J., and Aist, G. (2001). Evaluating tutors that listen. In (K. Forbus and P. Feltovich, Eds.) *Smart Machines in Education: The coming revolution in educational technology.* MIT/AAAI Press.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Platz, C., Sklar, M. B., and Tobin, B. (2001). A Controlled Evaluation of Computer- versus Human-assisted Oral Reading. Poster presented at the 10th International Conference on Artificial Intelligence in Education (AI-ED), 2001.

Mostow, J., Roth, S.F., Hauptmann, A. G., and Kane, M. (1994). A Prototype Reading Coach that Listens. *Proceedings of the Twelfth National Conference on Artificial Intelligence* (AAAI-94), Seattle WA, 1994. Selected as the AAAI-94 Outstanding Paper.

Nagy, W. E., Herman, P. A., and Anderson R. C. (1985). Learning Words from Context. *Reading Research Quarterly* 20(2), 233-253.

National Institute of Child Health and Human Development, Report of the National Reading Panel. (2000). Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction: Reports of the Subgroups (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office. Available online from http://www.nationalreadingpanel.org/

Robbins, C., and Ehri, L.C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. *Journal of Educational Psychology* 86(1), 54-64.

Sato, Satoshi. (2001). Automated Editing of Hypertext Resume from the World Wide Web. Proceedings of 2001 Symposium on Applications and the Internet (SAINT 2001), pp15-22, San Diego, California, 8-12 Janunary, 2001. Also: http://wit.kuee.kyoto-u.ac.jp/wit/eterm/

Schechter, J. B. (n.d.) The Reader: Interface and implementation. http://www.cogsci.princeton.edu/~wn/current/readerdoc.html

Scott, J. A., and Nagy, W. E. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly* 32(2), 184-200.