Factoids: Automatically constructing and administering vocabulary assistance and assessment

Greg Aist
aist@cs.cmu.edu
Project LISTEN, Robotics Institute,
Carnegie Mellon University, Pittsburgh PA 15213 USA
http://www.cs.cmu.edu/~listen

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.

1. Introduction

Project LISTEN's Reading Tutor listens to children read aloud, and helps them (http://www.cs.cmu.edu/~listen). The Reading Tutor displays a story (fiction or nonfiction) 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. The overall goal of the Reading Tutor is to help children learn to read. We focus here on one part of that overall goal: vocabulary acquisition – helping children learn the meaning of words.

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 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.

What kind of help should the Reading Tutor give? Options include:

- A conventional definition. "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 definition from a children's dictionary. 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 comparison to another word. "An astronaut is a kind of traveler."
- A short explanation. "An astronaut is someone who goes into outer space."
- An example sentence. "The astronaut went to the Moon in a rocket."

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 [1]. 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.

In the remainder of this paper we describe an experiment on vocabulary assistance. The experiment had two main components. First, automatically generating and presenting factoids – that is, comparisons to other words (drawn from WordNet; [2]). Second, automatically generating and administering assessments. We first give the five-step schema for the experiment, and then discuss each step in detail.

2. Experiment Design

The design of the experiment described in this paper 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.** Student reads story (with Reading Tutor assistance) up to just before the sentence containing the target word.
2. **Reading Tutor (sometimes) provides a factoid.**
   a. If this is a control trial, nothing extra happens
   b. If this is an experimental trial, the student reads a factoid (with Reading Tutor assistance).
3. **Student continues reading story.** Student reads remainder of the story, with Reading Tutor assistance.
4. **Time passes.** One or more days pass.
5. **Reading Tutor tests student's knowledge of the word.** The Reading Tutor administers a multiple-choice vocabulary question at the start of the session on the next day the student logs in.

We discuss each step in turn.

2.1. Student starts reading story

The Reading Tutor displayed the story one sentence at a time, and listened to the student read all or part of the sentence aloud. The Reading Tutor provided help in response to student mouse clicks and its analysis of the student's reading. Help available included reading all or part of the sentence aloud, sounding out a word either sound-by-sound or syllable-by-syllable, and

2.2. 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.

2.2.1. Selecting target words.

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 [3]. To construct a 4-item multiple-choice vocabulary question, the Reading Tutor needed the correct answer and three wrong answers to serve as distractors.

The Reading Tutor used synonyms and hypernyms as the correct answer, reverting to a sibling (Figure 1) only if neither a synonym nor a hypernym could be found.

The word must have at least three vocabulary question distractors. Vocabulary question distractors were cousins of the target word (words with a common grandparent but different parents) (Figure 1.) 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.

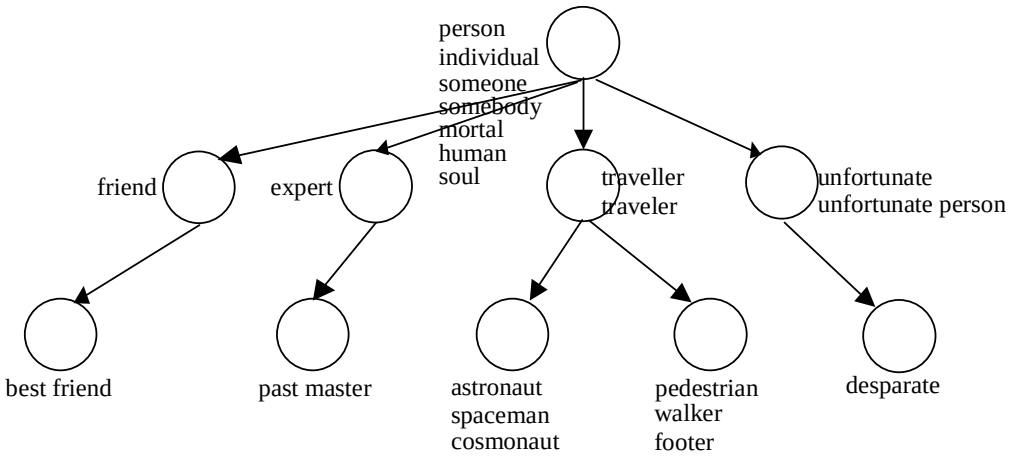

Figure 1. Siblings and cousins in WordNet. Selected portion of the WordNet 1.6 hierarchy. Nodes contain the set of all words that are synonyms of each other – that is, that form a single synset. Arrows point from general to more specific nodes. astronaut is a sibling to pedestrian because their nodes share the parent node "traveller; traveler". astronaut is a cousin to best friend because their nodes share the grandparent node "person; individual; … soul".

On the actual test, the answers were in a randomized 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.

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. [4], 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.

2.2.2. 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.

2.2.3. 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 [2]. WordNet, originally developed by George Miller and colleagues, 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.

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.)

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.)

2.3. 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.

2.4. 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.

2.5. 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.

We also needed to evaluate the effectiveness of vocabulary assistance. Nagy, Herman, and Anderson [3] 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.

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, to prevent bias and hamper cheating, as explained earlier. 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).

3. 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

		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.

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.

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.

4. 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 [5]. 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 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. Here, the outcome variable was whether the student got the answer right or not. The following factors were included in the model:

- whether 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.

4.1. 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.

4.2. Exploration revealed possible effect for low-frequency words with one sense, tested one or two days later

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).

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 [6]. Results were again not significant, but suggestive of a positive impact of factoids, as follows (Table 2).

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).

Table 2. Single-sense difficult words.

Number of trials	How were trials selected?	Per-student average number right	Coefficient in logistic regression model
720 trials	Single-sense words	34.9% ± 23.0% for control vs. 38.4% ± 26.5% for experimental	0.23 ± 0.17
191 trials	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	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	Single sense rare words	35.4% ± 30.5% for control vs. 42.4% ± 37.3% for experimental	.16 ± .29
189 trials	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

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

aluminum astronaut bliss bobbin coward crouching daisies eggshell glittering headdress hello infirmities liar outskirts pasta pebbles plat plumage pollen princess rwanda salad tennis twinkling vales wading wayside

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.

4.3. 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.

4.4. 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.

Table 4. Further characterization of 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 37.8% ± 22.3% expt.	No effect of factoid: -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
29 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
31 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

5. Lessons learned from factoid study

There were various problems with factoid assistance that may have diluted the effectiveness of factoids. We have already discussed target word frequency, multiple senses, and socially unacceptable factoids or test items. In addition, other problems remained. For the vocabulary assistance, some comparison words may have been harder to understand than the target words.

There were also various problems with the automated assessment that may have obscured the effectiveness of factoids. For example, some of the incorrect answers (distractors) were themselves rare – such as *butterfly pea* – making the question difficult to understand. Or, questions may have relied on uncommon knowledge, such a *banana* being (botanically) an herb.

In fact, at the end of Fall 1999, we turned off the vocabulary questions primarily because they were getting slower and slower as database queries labored over data collected during the entire year to date, but also due to the problems we just 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 [7], [8].

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

bear out [9]). 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

This work was supported in part by the National Science Foundation under Grant Nos. REC-9720348 and REC-9979894, and by Greg Aist's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

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). We also thank anonymous AI-ED reviewers for their comments, and Jack Mostow and Brian Tobin for reading and commenting on earlier drafts of this paper.

References

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

[2] C. Fellbaum, ed.., WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA, 1998.

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

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

[5] National Institute of Child Health and Human Development, Report of the National Reading Panel. 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, 2000. Available online from http://www.nationalreadingpanel.org/

[6] H. Kucera and W. N. Francis, Computational Analysis of Present-Day American English. Brown University Press, Providence, RI, 1967.

[7] J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, C. Platz, M. B. Sklar, B. Tobin, 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.

[8] G. Aist, P. Burkhead, A. Corbett, A. Cuneo, B. Junker, J. Mostow, M.B. Sklar, and B. Tobin, 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.

[9] G. Aist, Helping Children Learn Vocabulary During Computer-Assisted Oral Reading. Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2001.