Improving the Help Selection Policy in a Reading Tutor that Listens

Cecily Heiner, Joseph Beck, Jack Mostow
Project LISTEN, Carnegie Mellon University
RI-NSH, 5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213
{cecily, joseph.beck, mostow}@cs.cmu.edu

Abstract

What type of oral reading assistance is most effective for a given student on a given word? We analyze 189,039 randomized trials of a within-subject experiment to compare the effects of several types of help in the 2002-2003 version of Project LISTEN's Reading Tutor. The independent variable is the type of help given on a word. The outcome variable is the student's performance at the next encounter of that word, as measured by automatic speech recognition. Training a help selection policy sensitive to student or word level improves this outcome by a projected 4% – a substantial effect for picking a single better intervention.

1 Introduction

One aspect of tutoring (whether human or automated) is choosing what kind of help to give when a student needs assistance. The quality of these choices may affect the student's learning. This paper addresses the problem of choosing the best kind of help for a given situation.

We address this problem in the domain of children's oral reading. A number of studies have examined the effect of help type on children's reading, with sometimes conflicting findings. Some studies (e.g., Campbell, 1988) analyzed help given by teachers in classrooms. These studies were observational, limited to small samples, and sometimes conflated teacher with help type (e.g., Juel & Minden-Cupp, 1999). Some studies have evaluated computer-assisted reading (e.g., Reitsma, 1988), but few seem to have focused on help type. One such experiment (Spaai *et al.*, 1991) found an advantage for whole-word feedback over decomposing words into phonemes. Another experiment (Wise, 1992) found that it was less effective to decompose words into phonemes than into syllables, subsyllables, or not at all. Olson and Wise (1992) found that segmenting words into onset and rime was most helpful for less severely disabled subjects, but that syllable segmentation was best for more severely disabled subjects.

The present study exploits a novel research platform to expand the number of help types, the number of students, and the amount of data. Project LISTEN's Reading Tutor uses automatic speech recognition (ASR) to listen to children read aloud (Mostow & Aist, 2001; Mostow *et al.*, 2003). This paper is based on data from the 2002-2003 version, used daily on nearly 200 computers at nine elementary schools by hundreds of K-4 students spanning a wide range of reading proficiency.

The Reading Tutor embeds automated within-subject experiments (e.g., Aist, 2001) that use randomized trials to test alternative tutorial actions. By aggregating over many thousands of such trials and comparing effects of alternative choices on the ensuing tutorial dialogue, we can draw statistically reliable conclusions about their relative efficacy.

The Reading Tutor gives help at different levels, such as decoding a word, comprehending a sentence, or performing a task. This paper analyzes the efficacy of different types of word help. The Reading Tutor gives help on a word when it notices the student skip the word, misread the word, get stuck, or click for help (Mostow & Aist, 1999). In this paper, we do not distinguish between student initiated and tutor initiated help.

2 Experimental Design

The 2002-2003 version of the Reading Tutor randomized the selection of help type as an embedded experiment. The interface let students pick a particular type of help for a given word, but they seldom did, and we ignore this case here. Thus each experimental trial started with the Reading Tutor choosing which type of help to give on a particular word.

The randomized selection was controlled by a specified probability distribution, but constrained by feasibility considerations. For example, rhyming hints were infeasible for words like "orange." Other considerations excluded ineffective help such as sounding out very long words. The Reading Tutor selected among the following types of word help, listed here in decreasing order of frequency:

SayWord plays a recording of the word. For homographs, it gives both pronunciations.

WordinContext plays a recording of the word extracted from the sentence. The word must be at least three characters long.

Autophonics pronounces a selected grapheme. The word must be at least three letters long.

SoundOut plays video clips of a child's mouth saying the phonemes of the word. The word must be at least two characters long, not a homograph, and no longer than four phonemes.

Recue reads words in the sentence leading up to, but not including, the word. The word must be in the third position or later in the sentence.

OnsetRime says the first phoneme, pauses, and says the rest of the phonemes. The word must be at least three letters long and not a homograph.

StartsLike says "starts like (word with the same beginning)." The word must be two or more letters.

RhymesWith says "Rhymes with (rhyming word)." The words must be at least two letters long, and their rimes must be spelled the same.

Syllabify says the syllables of the word separated by short pauses. For want of syllable recordings, *Syllabify* approximates them by concatenating phoneme recordings. The word must be at least two characters long and cannot be a homograph.

ShowPicture shows a picture of the word.

SoundEffect plays a sound related to the word.

The outcome of a trial should test the efficacy of the help given. An earlier study (Aist & Mostow, 1998) used as trial outcome the student's next attempt to read the word while (re-)reading the same sentence. This outcome was appropriate for gauging immediate effects, but we want to know the effects on student learning. Accordingly, we consider the student's performance on the *next* encounter of the word in a later sentence, possibly on a later day. We define the outcome of each trial as whether the Reading Tutor's ASR accepted or rejected the word during the student's first utterance for this later sentence.

For example, a student clicks on the word "could." The Reading Tutor randomly chooses to give the rhyming hint "Rhymes with would." Later the student encounters the word "could" in a different sentence. The outcome of the trial is whether the Reading Tutor accepts or rejects the student's reading of "could" in the later sentence.

Sometimes the Reading Tutor gives help more than once on the same word during the same sentence, for example when the student clicks again. To simplify the analysis, we treat each such event as an independent trial with the same shared outcome, namely performance at the next encounter of the word.

3 Training a Better Help Policy

We measure the *efficacy* of a help type as the percentage of trials where the word was accepted at the next encounter. This percentage is somewhat low because we consider only the first utterance, so as to exclude effects of subsequent Reading Tutor assistance. This criterion penalizes children who take more than one utterance to read a sentence. Assuming this penalty is independent of the original help type, it may reduce the estimated *absolute* efficacy of different help types but should not distort their *relative* efficacy.

A *help selection policy* chooses which help to give. Section 2 described the randomized selection policy for the 2002-2003 version of the Reading Tutor. We measure the efficacy of this baseline policy as the overall average of all the types of help, weighted by how often each type was given.

What is the best help selection policy? One obvious answer is to pick whichever help type achieved the highest efficacy. However, we found that this approach suffered from poor estimates of efficacy for rare help types.

Instead, we select the help type with the highest confidence of outperforming the baseline. To handle our binary outcome variable and discriminate against poorly estimated values for rare help types, we adopt a Chi-Squared (χ^2) confidence measure:

$$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(a + c)(c + d)(b + d)}$$

Here a = the number of words accepted after the selected help type, b = the number of words rejected after the selected help type, c = the number of words accepted after a help type other than the selected help type, and d = the number of words rejected after a help type other than the selected help type.

Table 1: Help types, ordered by efficacy

Type of help	# times given	Efficacy ± std. error	χ2
RhymesWith	13,165	69.5 ± 0.4%	58.43
WordInContext	24,841	68.9 ± 0.3%	73.38
SoundEffect	488	68.6 ± 2.1%	1.14
ShowPicture	2,285	68.6 ± 1.0%	5.22
OnsetRime	14,223	68.3 ± 0.4%	23.52
StartsLike	13,671	67.2 ± 0.4%	4.08
SayWord	56,791	66.8 ± 0.2%	4.85
SoundOut	19,677	66.4 ± 0.3%	0.01
Overall	*189,039*	*66.4 ± 0.1%*	0.00
Autophonics	22,933	66.3 ± 0.3%	0.01
Syllabify	6,280	63.1 ± 0.6%	30.73
Recue	14,685	56.0 ± 0.4%	709.76

Table 1 lists types of help in decreasing order of efficacy with their χ^2 Values. WordInContext is the best help type even though its efficacy (68.9%) is lower than RhymesWith (69.5%). The reason is that WordInContext has the highest χ^2 (73.38) of help types that exceed the overall efficacy (66.4%). We measure the *improvement* of a help policy over the baseline as their difference in efficacy.

4 Evaluating the Training Method

Picking the best help type after the fact exploits information unavailable in advance and may overestimate efficacy by overfitting (Mitchell, 1997, p. 67). To evaluate our training method fairly, we use a twenty-fold cross validation, partitioning the students randomly into 20 disjoint sets. We train on 19 sets, test on the held-out set, repeat this procedure for each set, and average the results. Training simply picks the best help type *h* in a training set. Then we use the test set to compute the improvement achieved by always using *h*. To estimate the resulting improvement, we average the improvement for each test set, weighted by the amount of help given in that set. The result (1.9%) is the expected increase in efficacy for unseen students drawn from a similar distribution.

Although we want to know which help type was most helpful overall, we are more interested in building a help policy conditioned on the student and the words. For this study, we use a simple analysis based on four grade levels of student proficiency and three grade levels of word difficulty. We measure student proficiency using grade-equivalent Woodcock Reading Mastery Test (Woodcock, 1998) Word Identification pre-test scores. Proficiencies range from 0.5 to 9.9, with a mean of 1.9. For simplicity, we round to the nearest integer and group all students at level 4 and above together. We use a heuristic to estimate a grade-equivalent level for each word. Word levels range from 0.6 to 11.2, with a mean of 2.3. Again for simplicity, we round to the nearest integer and group together all words at and above grade 3.

To train a help selection policy conditioned on students, words, or both, we disaggregate the data into subsets by student proficiency, word level, or both. We cross-validate the training method within each subset of a disaggregation, e.g., each grade level. We average the resulting improvement across the subsets, weighted by the amount of help given in each subset. Conditioning on student proficiency improves help selection by 3.9% over the baseline. Conditioning on word level improves help selection by 3.7% over the baseline. Conditioning by both improves only 3.1% over the baseline.

What helped which students or words the most? Table 2 shows the type(s) of help rated best for each level of student reading proficiency, how many of the 20 training sets rated it best, and the average increase in efficacy over the baseline for that level. Table 3 shows a similar breakdown for word level.

We include the number of training sets to indicate the closeness of the contest for the best help type. A help type rated best in all 20 training sets is likely to be genuinely better than other types of help. If another help type is almost as good, one might expect it to be rated best in at least one of the 20 training sets.

Table 2 shows that training a help policy has the greatest impact for students at a grade 2-3 level, with RhymesWith as the most effective help type for these readers. The number of training sets at grade level 4 is only 18 because the random partitioning of students into test sets led to two test sets containing no students at that level.

Table 2: Best help types, by student level

Student level	Best help type(s)	Efficacy increase
1	WordInContext (20/20)	2.6%
2	RhymesWith (20/20)	4.8%
3	RhymesWith (20/20)	5.8%
4	WordInContext (17/18) StartsLike (1/18)	0.2%

Table 3 shows that training a help policy has the greatest impact for words at a grade 1 level. Rhyming hints (RhymesWith and OnsetRime) were best for these words. Whole-word help (WordInContext and SayWord) was more effective for harder words. Perhaps the level-varying preferences for SayWord versus WordInContext involve differences in intelligibility or in how they treat homographs – giving both pronunciations, or only the right pronunciation for the current context.

Table 3: Best help types, by word difficulty

Word difficulty	Best help type(s)	Efficacy increase
1	OnsetRime (18/20) RhymesWith (2/20)	5.0%
2	WordInContext (20/20)	3.2%
3	SayWord (20/20)	3.4%

5 Conclusion and Future Work

What type of help is most effective for a tutor to give? We study this question in the context of an automated tutor for children's oral reading.

We show how an automated tutor that listens can collect ecologically valid, fine-grained measures of learning in larger quantities than would otherwise have been feasible. Conducting, logging, aggregating, and analyzing large numbers of randomized trials can overcome the limitations of ASR well enough to reveal subtle differences in efficacy of tutorial actions, and their interactions with student and word variables. Future work may contrast student- vs. tutor-initiated help, single vs. multiple help, same- vs. later-day encounters, and group characteristics vs. individual differences.

We quantify the efficacy of several help types. Rhyming hints worked best for grade 2-3 level readers and easy words, while whole-word help worked best for grade 1 and 4 level readers and harder words. Recue performed worst of all the help types. This intervention might help students identify a word in a context where the word is predictable, but apparently such contexts were rare, or such scaffolding did not help the student identify the word in a later context. These results are for specific implementations of the help types, and might differ for other implementations.

Without disaggregation, the best help type improves by 1.9% over the baseline help policy. Disaggregating by student or word level doubles this expected improvement. Disaggregating by both seems to cause overfitting.

We estimate the effects of a help policy by selecting the single best help type in a given situation. This estimate assumes the efficacy of different help events is independent. However, variety might make help more effective. The effectiveness of one help type might depend on knowledge gained from another help type (Mostow & Aist, 2001, p. 197). On the other hand, varied help types might confuse young readers, in which case a single help type would work better. To evaluate the true efficacy of trained help policies, we must test them in the Reading Tutor – and link them to longer-term student learning gains.

Acknowledgements

This work was supported by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and 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.

References

Aist, G. (2001). Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education, 12*, 212-231.

Aist, G., & Mostow, J. (1998, March). Estimating the effectiveness of conversational behaviors in a reading tutor that listens. *Working Notes of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Stanford, CA.

Campbell, R. (1988). *Hearing Children Read.* London and New York: Routledge.

Juel, C., & Minden-Cupp, C. (1999). *Learning to Read Words: Linguistic Units and Strategies*. Ann Arbor, MI: CIERA.

Mitchell, T. (1997). *Machine Learning*. New York: McGraw-Hill.

Mostow, J., & 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.

---. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B., & Tobin, B. (2003). Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*(1), 61-117.

Olson, R. K., & Wise, B. W. (1992). Reading on the computer with orthographic and speech feedback. *Reading & Writing Interdisciplinary Journal, 4*, 107-144.

Reitsma, P. (1988). Reading Practice for Beginners: Effects of Guided Reading, Reading-while-Listening, and Independent Reading with Computer-Based Speech Feedback. *Reading Research Quarterly, 23*(2), 219-235.

Spaai, G. W., Ellermann, H. H., & Reitsma, P. (1991). Effects of segmented and whole-word sound feedback on learning to read single words. *Journal of Educational Research, 84*(4), 204-213.

Wise, B. W. (1992). Whole Words and Decoding for Short-Term Learning: Comparisons on a "Talking-Computer" System. *Journal of Experimental Child Psychology, 54*(2), 147-167.

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. Circle Pines, Minnesota: American Guidance Service.