Using knowledge tracing to measure student reading proficiencies

Joseph E. Beck[1] and June Sison[2]

{joseph.beck, sison}@cs.cmu.edu
Phone: +1 412 268 5726
[1] Center for Automated Learning and Discovery
[2]Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213. USA.

Abstract. Constructing a student model for language tutors is a challenging task. This paper describes using knowledge tracing to construct a student model of reading proficiency and validates the model. We use speech recognition to assess a student's reading proficiency at a subword level, even though the speech recognizer output is at the level of words. Specifically, we estimate the student's knowledge of 80 letter to sound mappings, such as *ch* making the sound /K/ in "chemistry." At a coarse level, the student model did a better job at estimating reading proficiency for 47.2% of the students than did a standardized test designed for the task. Our model's estimate of the student's knowledge on individual letter to sound mappings is a significant predictor of whether he will ask for help on a particular word. Thus, our student model is able to describe student performance both at a coarse- and at a fine-grain size.

1 Introduction

Project LISTEN's Reading Tutor [8] is an intelligent tutor that listens to students read aloud with the goal of helping them learn how to read English. Target users are students in first through fourth grades (approximately 6- through 9-year olds). Students are shown one sentence (or fragment) at a time, and the Reading Tutor uses speech recognition technology to (try to) determine which words the student has read incorrectly. Much of the Reading Tutor's power comes from allowing children to request help and from detecting some mistakes that students make while reading. It does not have the strong reasoning about the user that distinguishes a classic intelligent tutoring system, although it does base some decisions, such as picking a story at an appropriate level of challenge, on the student's reading proficiency.

We have constructed models that assess a student's overall reading proficiency [2], but have not built a model of the student's performance on various skills in reading. Much of the difficulty comes from the inaccuracies inherent in speech recognition. Providing explicit feedback based only on student performance on one attempt at reading a word is not viable since the accuracy at distinguishing correct from incorrect reading is not high enough [13]. Due to such problems, student modeling has not received as much attention in computer assisted language learning systems as in classic ITS [5], although there are exceptions such as [7].

Our goal is to use speech recognition to reason about students' proficiency at a finer grain-size. Even if it is not possible to provide immediate feedback for student mistakes, it may be possible to collect enough data over time to estimate a student's proficiency at various aspects of reading. Such a result would be helpful for other tutors that use speech input, particularly language tutors. Our approach is to use knowledge tracing to assess student reading skills.

2 Knowledge tracing

Knowledge tracing [4] is an approach for estimating the probability a student knows a skill given observations of him attempting to perform the skill. First we briefly discuss the parameters used in knowledge tracing, then we describe how to modify the approach to work with speech recognition.

2.1 Parameters in knowledge tracing

For each skill in the curriculum, there is a P(k) representing the probability the student knows the skill, and there are also two learning parameters:

- P(L0) is the initial probability a student knows a skill
- P(t) is the probability a student learns a skill given an opportunity

However, student performance is a noisy reflection of his underlying knowledge. Therefore, there are two performance parameters for each skill:

- P(slip) = P(incorrect | know skill), i.e., the probability a student gives an incorrect response even if he has mastered the skill. For example, hastily typing "32" instead of "23."
- P(guess) = P(correct | didn't know skill), i.e. the probability a student manages to generate a correct response even if he has not mastered the skill. For example, a student has a 50% chance of getting a true/false question correct.

When the tutor observes a student respond to a question either correctly or incorrectly, it uses the appropriate skill's performance parameters (to discount guesses and slips) to update its estimate of the student's knowledge. A fuller discussion of knowledge tracing is available in [4].

2.2 Accounting for speech recognizer inaccuracies

Although knowledge tracing updates its estimate of the student's internal knowledge on the basis of observable actions, this approach is problematic with the Reading Tutor since the output of automated speech recognition (ASR) is far from trustworthy. Figure 1 shows how both student and interface characteristics mediate student performance. In standard knowledge tracing, there is no need for the intermediate nodes or their transitions to the observed student performance. However, since our observations of the student are noisy, we need additional possible transitions. FA stands for the probability of a False Alarm and MD stands for the probability of Miscue Detection. A false alarm is when the student reads a word correctly but the word is rejected by the ASR; a detected miscue is when the student misreads a word and it is scored as incorrect by the ASR. In a perfect environment, FA would be 0 and MD would be 1, and there would therefore be no need for the additional transitions. Overall in the Reading Tutor, FA ≈ 0.04 and MD ≈ 0.25 (only counting cases where the student said some other word, the tutor is much better at scoring silence as incorrectly reading a word).

All we are able to observe is whether the student's response is scored as being correct, and the tutor's estimate of his knowledge. Given these limitations, any path that takes the student from knowing a skill to generating an incorrect response is considered a slip; it does not matter if the student actually slipped, or if his response was observed as incorrect due to a false alarm. Similarly, a guess is any path from the student not knowing the skill to an observed correct performance. Therefore, can define two additional variables slip' and guess' to account for both paths:

slip' = slip * MD + (1-slip) * FA

guess' = guess * (1-FA) + (1-guess) * (1-MD)

Since we expect ASR performance to vary based on the words being read, it is not appropriate to use a constant MD and FA for all words. Therefore, when we observe a slip, while it would be informative to know whether it was caused by the student or the ASR, there is no good way of knowing which is at fault. As a result, we do not try to infer the FA, MD, slip, and guess parameters. Instead, we directly estimate the slip' and guess' parameters for each skill directly from data (see Section 3.4).

For simplicity, we henceforth refer to guess' and slip' and guess and slip. However, note that the semantics of P(slip) and P(guess) change when using knowledge tracing in this manner. These parameters now model both the student and the method for scoring the student's performance. However, the application of knowledge tracing and the updating of student knowledge remain unchanged.

3 Method for applying knowledge tracing

We now describe how we applied knowledge tracing to our data. First we describe the data collected, next we describe the reading skills we modeled, then we describe how to determine which words the student attempted to read, and finally discuss the knowledge tracing parameter estimates.

3.1 Description of data

Our data come from 284 students who used the Reading Tutor in the 2002-2003 school year. The students using the Reading Tutor were part of a controlled study of learning gains, so were pre- and post-tested on several reading tests. Students were administered the Woodcock Reading Mastery Test [14], the Test of Written Spelling

[6], the Gray Oral Reading Test [12], and the Test of Word Reading Efficiency [11]. All of these tests are human administered and scored.

Students' usage ranged from 27 seconds to 29 hours, with a mean of 8.6 hours and a median of 5.9 hours. The 27 seconds of usage was anomalous, as only four other users had less than one hour of usage.

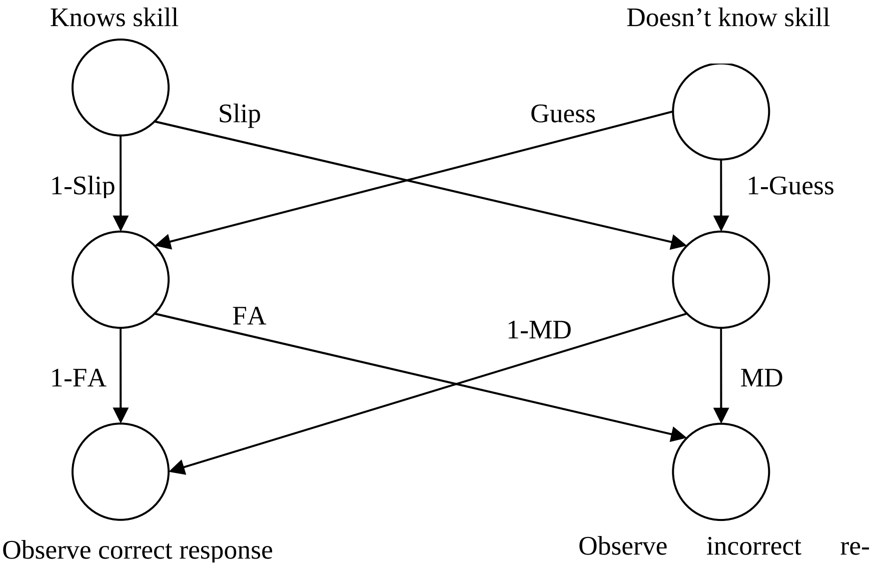

Figure 1. Knowledge tracing with imperfect scoring of student responses.

While using the Reading Tutor, students read from 3 words to 35102. The mean number of words read was 8129 and the median was 5715. When students read a sentence, their speech was processed by the ASR and aligned against the sentence [10]. This alignment scores each word of the sentence as either being accepted (heard by the ASR as read correctly), rejected (the ASR heard and aligned some other word), or skipped. In Table 1, the student was supposed to read "The dog ran behind the house." The bottom row of the table shows how the student's performance would be scored by the tutor.

Table 1. Example alignment of ASR output to sentence

Sentence	The	dog	ran	behind	the…
ASR output	The	the	ran		
Scoring	Accept	Reject	Accept	Skipped	Skipped

3.2 What reading skills to assess?

Given the ASR's judgment of the student's reading, we must decide which reading skills we wish to assess. We could measure the student's competency on each word in the English language, but such a model would suffer from sparse data problems and would not generalize to new words the student encounters. Instead, we assess a student's knowledge of grapheme→phoneme (g→p) mappings. A grapheme is a group of letters in a word that produces a particular phoneme (sound). So our goal is to assess the student's knowledge these g→p mappings.

For example, *ch* can make the /CH/ sound as in the word "Charles." However, *ch* can also make the /K/ sound as in "chaos." By assessing students on the component skills necessary to read a word, we hope to build a model that will allow the tutor to make predictions about words the student has not yet seen. For example, if the student cannot read "chaos" then he probably cannot read "chemistry" either.

Modeling the student's proficiency at a subword level is difficult, as we do not have observations of the student attempting to read g→p mappings in isolation. There are two reasons for this lack. First, speech recognition is imperfect differentiating individual phonemes. Second, the primary goal of the 2002-2003 Reading Tutor is to have students learn to read by reading connected text, not to read isolated graphemes with the goal of allowing the tutor to assess their skills. To overcome this problem, we apply knowledge tracing to the individual g→p mappings that make up the particular word. For example, the word "chemist" contains ch→/K/, e→/EH/, m→/M/, i→/IH/, s→/S/, and t→/T/ as g→p mappings.

However, which mappings are indicative of a student's skill? Prior research on children's reading [9] shows that children are often able to decode the beginning and end of a word, but have problems with the interior. Therefore, we ignore the first and last g→p mappings of a word and use the student's performance reading to word to update the tutor's estimate of the student's knowledge of the interior g→p mappings. In the above example we would update the student's knowledge on e→/EH/, m→/M/, i→/IH/, and s→/S/. Words with fewer than three graphemes do not adjust the estimate of the student's knowledge.

3.3 Which words to score?

When students read a sentence in the Reading Tutor, sometimes they do not attempt to read all of the words in the sentence. If the student pauses in his reading, the ASR will score what the student has read so far. For example, in Table 1, the student appears to have gotten stuck on the word "behind" and stopped reading. It is reasonable to infer the student could not read the word "behind." However, the scoring of "the" and "house" depends on what skills are being assessed. If the goal is to measure the student's overall reading competency, then counting those words as read incorrectly will provide a better estimate since stronger readers will need to pause fewer times. Informal experiments on our data bear out this idea.

However, our goal is not to assess a student's overall reading proficiency, but to estimate his proficiency at particular g→p mappings. For this goal, the words "the" and "house" provide no information about the student's competency on the mappings that make up those words. Therefore we do not apply knowledge tracing to those words.

More formally, we estimate the words a student attempted as follows:

1. *i* = Find the first word in the sentence that was accepted
2. *j* = Last word in the sentence that was accepted
3. Apply knowledge tracing to sentence words *i…j+1*

In the example in Table 1, *i=1* and *j=3*, and the words 1-4 would be scored ("The dog ran behind"). This heuristic assumes the reason the student stopped reading was because he could not read the next word in the sentence.

3.4 Parameter estimation

We have described how to take the aligned ASR output and to use a heuristic to determine which words to score, and which g→p mappings in the words to model. The next step is to estimate the four knowledge tracing parameters (L0, t, guess, slip) for the set of data collected from students.

There are 429 distinct g→p mappings that occur in at least one word in our dictionary. We used the student's performance on words containing those mappings as input to an optimization algorithm[1] to fit the four knowledge tracing parameters for each g→p using our students' performance data. We then restricted the set of mappings to those with at least 1000 attempts combined from all students. We also removed mappings that fit the knowledge tracing model poorly; we required an R^2 of 0.20. These restrictions limited the set to 80 mappings.

The optimization code required some modification since it was designed for more traditional knowledge tracing. For example, the code restricted the number of "exercises" where students get to apply a particular skill to be less than 100. In our case, an exercise is a student attempting to read a word containing a particular g→p mapping. Some students encounter a particular mapping thousands of times. Another restriction is that P(guess) was forced to be less than 0.3 and P(slip) to be less than 0.1. For our task, such a restriction is inappropriate as mappings with at least 10,000 observations had an average P(guess) of 0.71 and P(slip) of 0.13.

The reason P(guess) is so high is that the Reading Tutor is biased towards hearing the student read the sentence correctly in order to reduce frustration from novices having correct reading scored as incorrect. These data demonstrate that with current speech recognition technology, a tutor cannot provide the same type of immediate feedback as a tutor with typed input due to the uncertainty in whether the student was correct. With such a high guess parameter, many observations are required for a student to be considered proficient in a skill. Fortunately, students read hundreds of words each day they use the Reading Tutor, so the bandwidth should be sufficient to estimate the student's proficiencies.

Once the above steps have been performed, we have a set of knowledge tracing parameter estimates for 80 g→p mappings. By taking the aligned output of the ASR of the student's reading, we can apply the knowledge tracing model to estimate the student's proficiency on each skill. This process results in a probability estimate as to whether the student knows each of the 80 reading skills in our model.

[1] Source code is courtesy of Albert Corbett and Ryan Baker and is available at http://www.cs.cmu.edu/~rsbaker/curvefit.tar.gz

4 Validation

We now discuss validating our model of the student's reading proficiency. First we demonstrate that, overall, it is a good model of how well students can identify words. Then we show that the individual g→p estimates have predictive power.

4.1 Performance at predicting Word ID scores

If we run knowledge tracing over the student's Reading Tutor performance for the year, we get a set of 80 probabilities that estimate the student's proficiency at each g→p mapping. To validate the accuracy of these probabilities, we use them to predict the student's Word Identification (WI) post test score from the Woodcock Reading Mastery Test [14]. The posttest occurred near the end of the school year. For the WI test, a human presents words for the student to read and records whether the student read the word correctly or not, and terminates the test when the student gets four words in a row incorrect. The WI test is a good test for validating the overall accuracy of our g→p mappings since it presents students with a series of words; the student then either recognizes the word on sight or segments the words into graphemes and produces the appropriate phonemes.

The goal is to use the estimates of the student's knowledge of the 80 g→p mappings to predict his grade equivalent WI post test score. Grade equivalent scores are of the form grade.month, for example 3.4 corresponds to a third grader in the fourth month of school. The month portion range from 0 to 9, with summer months excluded.

Grade equivalent can be misleading. For example, a math test of simple addition may show that a first-grader had a score of 5.3. This result does **not** mean the student has the math proficiency of a fifth grader, rather it means that he scored as well as a fifth grader might be expected to do on that test (so the student is quite skilled at addition, but the score says nothing about his knowledge of other math skills a fifth grader would be expected to know, such as fractions).

In contrast, many reading tests are designed for grades K-12 (roughly ages 5 through 17). For example, in WI, the test starts with easy words such as "red" and "the." For a student to receive a score of 5.3, the student would have to read words such as "temporal" or "gruffly." If a first grader can read such words (and the preceding words on the test), it is not unreasonable to say he can identify words as well as a fifth grader (although his other reading skills may be lacking). As a target for building a model of the student, the grade equivalent scale is a reasonable choice due to its interpretability by researchers. This use of grade equivalent scores follows guidelines [1] for when their use is appropriate.

We expect different g→p mappings to be predictive for students in different grades since skills that students have mastered in prior grades are unlikely to remain predictive in later grades. Therefore, we constructed a model for each grade. We entered terms into the regression model until the change in R^2 was less than 0.01 for grades one and two and less than 0.05 for grades three and four (there were fewer students in grades 3 and 4). This process resulted in ten mappings entering the model for grade one, 25 mappings for grade two, five mappings for grade three, and four mappings for grade four.

The resulting regression model for WI scores had, using a leave-one-out cross validation, an overall correlation of 0.88 with the WI test. It is reasonable to conclude that our model of students' word identification abilities is in reasonable agreement with a well-validated instrument for measuring the skill. We examined the case where our model's error from the student's actual WI was greatest: a fourth grader whose pretest WI score was 3.9, her posttest was 3.3, and our model's prediction was 6.1. It is unlikely the student's proficiency declined by 0.6 grade levels over the course of the year, and it was unclear whether we should believe the 3.3 or the 6.1. Perhaps our model is more trustworthy than the gold standard against which we validated it? There are a variety of reasons not to trust a single test measurement, including that it was administered on a particular day. Perhaps the student was feeling ill or did not take the test seriously? Also, we would like to know if our measure is better than WI. To get around these limitations, we looked at an alternate method of measuring word identification.

4.2 Alternate measure of word identification

To find an alternate method of measuring word identification, we examined our battery of tests we administer to students to find a set of tests that are most similar to WI:

1. The Accuracy score from the Gray Oral Reading Test (GORT) measures how many mistakes students make reading connected text. It correlates with WI at 0.76.
2. Sight Word Efficiency (SWE) from Test Of Word Reading Efficiency (TOWRE) measures how quickly students can decode common words. It correlates with WI at 0.80.
3. The Test of Written Spellling (TWS) is the opposite of word identification as students are presented a sound and asked to generate the proper letters, but is related to word identification [3]. It correlates with WI at 0.86.

None of these measures perfectly matches the construct of word identification, but they measure closely related constructs. We took the mean of these three tests as a proxy for the students' word identification proficiency. These tests were sometimes administered on different days and usually by different testers. The mean of the three tests correlates with WI at 0.87. Furthermore, the mean of the 3 scores (hereafter called WI3) does not suffer nearly as badly as WI from students dropping several months in proficiency from pre- to post-test. Given the stability of the WI3 measure, its being composed of constructs closely related to word identification, and its statistical correlation with WI, we feel it is a good measure of the students "true' word identification score.

Returning to the student whose WI posttest score deviated from the model. Her WI score was 3.3 her predicted score was 6.1, and her WI3 score was 5.1. Perhaps our model did a better job for this student than the WI test? To evaluate the accuracy of our model, we compared our model and the WI score to see how often each was closer to the WI3 score. The WI test was closer to the WI3 score 52.8% of the time, while our model was closer 47.2% of the time. An alternate evaluation is to examine the mean absolute error (MAE) between each estimate and WI3. WI had an MAE of 0.71 (SD of 0.56), while our model had an MAE of 0.77 (SD of 0.67), a difference of only 0.06 GE (roughly three weeks). So our model was marginally worse than the WI test at assessing (a proxy for) a student's word identification abilities. However, the WI test is a well-validated instrument, and to come with 0.06 GE of it is an accomplishment. Although marginally worse than the paper test, the knowledge tracing model can estimate the student's proficiency at any time throughoutthe school year, and requires no student time to generate an assessment.

4.3 Predicting help requests

To validate whether our model's estimates of the student's knowledge of individual g→p mappings were accurate, we predicted whether the student would ask for help on a word. We used help requests rather than the student's performance at reading words since we already extracted considerable data about student reading performance to build our model. Thus using it to confirm our model would be circular.

To measure whether knowledge of g→p mappings would help predict whether the student would ask for help, we examined every word the student encountered and noted whether he asked for help or not. We excluded words composed of fewer than three graphemes (since our model is based on student performance on interior g→p mappings). Approximately 79% of English tokens in children's reading materials are composed of 3 or more graphemes. The above restrictions limited us to 288,614 sentence word tokens the students encountered during their time using the Reading Tutor.

We constructed a logistic regression model to predict whether a student would ask for help on a word. This model had several components:

1. The identity of the student was a factor. Adding the student to the model controls for overall student ability, student differences in help-seeking behavior (in the past, student help request rates have differed by two orders of magnitude in the Reading Tutor).
2. The difficulty of the word (on a grade equivalent scale) was a covariate. Presumably students are more likely to ask for help on difficult words.
3. The position of the word in the sentence was a covariate. In the Reading tutor, students sometimes do not read the entire sentence. Therefore, we suspected that words earlier in the sentence are more likely to be clicked on for help.
4. The average knowledge of the 80 g→p for the student at the point in time when he encountered the word was a covariate. This term modeled the changes in the student's knowledge over the course of the year.
5. The student's average knowledge of the g→p mappings that composed the word, excluding the first and last mappings. For our data, of words with 3 or more graphemes, the modal number of graphemes was 3 and the median was 4. Therefore, there are generally only one or two interior

g→p mappings, so the student's average knowledge of the mappings in a word was not a broad description of the student's competencies, but is a focused description of his knowledge of the components of this word.

Logistic regression generates Beta coefficients to determine each variable's influence on the outcome. The Beta coefficients were 0.48 for word difficulty, -0.96 for the student's overall ability, -0.38 for the student's mean proficiency of the g→p mappings in the word, and -0.035 for the word's position in the sentence. If a variable has a positive Beta coefficient, then as the variable's value increases the student's probability of asking for help increases. Conversely, a negative Beta implies as the value increases, the student's probability of requesting help decreases. All of the Beta values were significant at P<0.001, and all point in the intuitive direction: as students become more proficient at reading they ask for help less, if a student has a higher estimated knowledge of the g→p mappings in this particular word, even after controlling for word difficulty, then the student is less likely to ask for help. It is important to note that the Beta values are not normalized in logistic regression, thus it is not appropriate to order the various features by how much predictive power they have.

These results provide evidence that individual estimates of the student's proficiency on g→p mappings are meaningful indicators of proficiency.

5 Conclusions and Future Work

This paper demonstrates that it is possible to apply classic student modeling techniques to language learning tutors that use speech recognition. While it is true the data are extremely noisy, it is possible to account for the noise and model student proficiency on subword skills, in our case g→p mappings, of reading. This model of proficiency is accurate in the aggregate since it is able to assess a student's word identification proficiency nearly as well as a paper test designed for the task. Furthermore, the individual estimates of the student's knowledge are also useful, since they predict whether a student requests help on a word.

Next steps for this work include a better model of credit assignment for words that are accepted or rejected. If the ASR believes the student made a mistake, it may not be fair to blame all of the interior g→p mappings, the blame should be spread probabilistically. Similarly, a student may generate correct reading without knowing all of the g→p mappings in a word.

Similarly, we will investigate a better model of how children decode words. For example, although early readers tend to understand the first part of a word, students who are just starting to read may struggle at this step. A model of children's reading that treats each component of the word as a separate skill would account for this problem.

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. We also acknowledge members of Project LISTEN who contributed to the design and development of the Reading Tutor, and the students who used the tutor.

References

1. *Canadian Psychological Association: Guidelines for Educational and Psychological Testing*. 1996: Also available at: http://www.acposb.on.ca/test.htm.
2. Beck, J.E., P. Jia, and J. Mostow. *Assessing Student Proficiency in a Reading Tutor that Listens*. in *Ninth International Conference on User Modeling*. 2003.p. 323-327 Johnstown, PA.
3. Carver, R.P., *The highly lawful relationship among pseudoword decoding, word identification, spelling, listening, and reading.* Scientific Studies of Reading, 2003. **7**(2): p. 127-154.
4. Corbett, A. and J. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge.* User modeling and user-adapted interaction, 1995. **4**: p. 253-278.
5. Heift, T. and M. Schulze, *Student Modeling and ab initio Language Learning.* System, the International Journal of Educational Technology and Language Learning Systems, 2003. **31**(4): p. 519-535.
6. Larsen, S.C., D.D. Hammill, and L.C. Moats, *Test of Written Spelling*. fourth ed. 1999, Austin, Texas: Pro-Ed.
7. Michaud, L.N., K.F. McCoy, and L.A. Stark. *Modeling the Acquisition of English: an Intelligent CALL Approach"*. in *Eighth International Conference on User Modeling*. 2001.p.: Springer-Verlag.
8. Mostow, J. and G. Aist, *Evaluating tutors that listen: An overview of Project LISTEN*, in *Smart Machines in Education*, K. Forbus and P. Feltovich, Editors. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.
9. Perfetti, C.A., *The representation problem in reading acquisition*, in *Reading Acquisition*, P.B. Gough, L.C. Ehri, and R. Treiman, Editors. 1992, Lawrence Erlbaum: Hillsdale, NJ. p. 145-174.
10. Tam, Y.-C., J. Beck, J. Mostow, and S. Banerjee. *Training a Confidence Measure for a Reading Tutor that Listens*. in *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. 2003.p. 3161-3164 Geneva, Switzerland.
11. Torgesen, J.K., R.K. Wagner, and C.A. Rashotte, *TOWRE: Test of Word Reading Efficiency*. 1999, Austin: Pro-Ed.
12. Wiederholt, J.L. and B.R. Bryant, *Gray Oral Reading Tests*. 3rd ed. 1992, Austin, TX: Pro-Ed.
13. Williams, S.M., D. Nix, and P. Fairweather. *Using Speech Recognition Technology to Enhance Literacy Instruction for Emerging Readers*. in *Fourth International Conference of the Learning Sciences*. 2000.p. 115-120: Erlbaum.
14. Woodcock, R.W., *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.