Automatic Natural Language Processing and the Detection of Reading Skills and Reading Comprehension

Chutima Boonthum-Denecke[a], Philip M. McCarthy,[b] Travis A. Lamkin,[b] G. Tanner Jackson[b], Joseph P. Magliano[c], and Danielle S. McNamara[b]

[a]Hampton University, Hampton VA, 23668, USA
[b]University of Memphis, Memphis, TN, 38152, USA
[c]Northern Illinois University, DeKalb, IL, 60115, USA
{chutima.boonthum, philmccarthy1, travis.lamkin, gtannerjackson, joemagliano, dsmcnamara1}@gmail.com,

Abstract

The primary goal of this study is to assess two approaches for detecting comprehension processes in R-SAT (Reading Strategy Assessment Tool). One approach is based on Latent Semantic Analysis (LSA) while the other is a combination of literal word matching and soundex. A secondary goal is to assess the potential for detecting specific reading comprehension strategies, either in isolation or combination. Participants typed "think-aloud" protocols while reading texts presented on computers. Human judges rated these protocols for the presence of the various reading comprehension strategies. LSA, word, and combined algorithms were compared and the results showed that a combination of both approaches yielded the best results. However, performance of the combined algorithm varied in terms of the type of processes and the grain size of the human coding system. Lastly, the use of reading strategies (either in isolation or combination) is positivity related to students' Gates–MacGinitie reading comprehension scores, which illustrates the merit of this approach for assessing comprehension skill.

Introduction

Reading Strategy Assessment Tool (R-SAT; Magliano, Millis, The R-SAT Development Team, Levinstein, and Boonthum, in press) is a computerized assessment that contains algorithms designed to identify students' reading strategies and level of comprehension *as* they read. Most other approaches attempt to assess strategies and comprehension *after* reading and often via multiple-choice tests. Assessments in R-SAT are based on typed verbal protocols that resemble think aloud protocols. To analyze these protocols, they are compared to "semantic benchmarks" that reflect comprehension processes. Two natural language processing (NLP) algorithms were considered: Latent Semantic Analysis (LSA) and word matching (literal match and soundex match). Both algorithms have been previously evaluated within iSTART (Interactive Strategy Training for Active Reading and Thinking; McNamara et al. 2009). However, the goal of assessment in iSTART is to provide a general assessment on the quality of a student's self-explanation. In contrast, R-SAT assessments are designed to detect specific strategies, such as *paraphrasing*, *bridging*, and *elaboration*. Although we have had reasonable success with word-matching algorithms (Magliano et al. in press), it is important to evaluate whether other approaches such as LSA (Landauer et al. 2007) can improve performance.

Reading Strategy Assessment Tool (R-SAT)

In R-SAT (Magliano et al. in press), a text is presented to the readers one sentence at a time. At specified target sentences, the readers are asked to either type in their

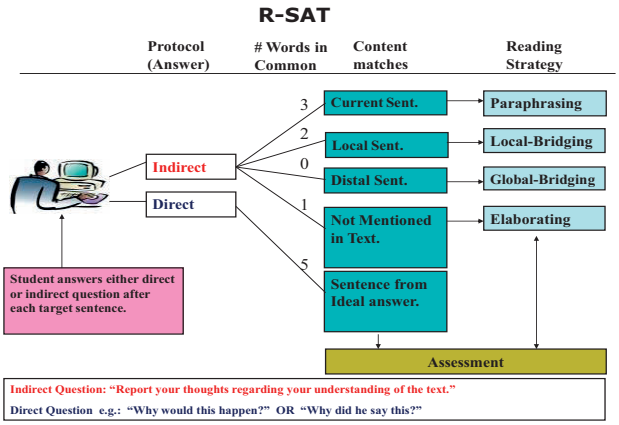

Figure 1. R-SAT System (Gilliam et al., 2007)

thoughts/understanding about the text or answer a *wh*-question (e.g., *why* or *how* question). The current study focuses on the typed response protocols, which will be evaluated with word matching algorithms alone, LSA alone, and a combination of the two.

Benchmarks. The response protocols from R-SAT are compared against strategy *benchmarks*, which are a set of words or phrases that represent each strategy. The benchmarks are defined as follows:

- Paraphrasing (P) - relation to the current sentence
- Local-Bridging (L) - relation to the immediate prior sentence in the text
- Distal-Bridging (D) - relation to all prior sentences in the text, excluding the immediate prior
- Elaboration (E) - relation to all subsequent sentences in the text, and words that are not present in the text.

Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA; Landauer 2007) is a high dimensional linear associative method that uses a statistical model for determining semantic similarity. This statistical approach requires a knowledge base constructed from a large corpus of related texts. Meanings are represented in terms of their similarity to other words from the corpora of documents. This approach uses induction-dimension optimization that greatly increases its learning ability for inferring indirect similarity relations. Although LSA has some well-established drawbacks, such as word order, alternate word forms, or potential misspellings, the approach has been successfully used to detect comprehension processes (e.g., Magliano and Millis 2003).

R-SAT uses LSA to compute the similarity between the protocol and benchmarks (LSA cosine) with the following equation:

$$Sim(D1, D2) = \frac{\sum_{i=1}^{d}(D1_i X\ D2_i)}{\sum_{i=1}^{d}(D1_i)^2\ X\ \sum_{i=1}^{d}(D2_i)^2}$$

Where d is a number of reduced dimensions in LSA matrix, D1 is a vector of a benchmark and D2 is a vector of a reader's protocol. Each document vector (D) contains eigenvalues indicating how a document D is in relation with the matrix space.

Word Matching (WM)

Word matching is one of the computationally lightest ways to evaluate natural language. A matching count can be computed by comparing words from two different sources against one another. Word matching can be performed in at least two ways: (1) Literal word matching and (2) Soundex matching

Literal word matching is the process of evaluating the similarity of two words by comparing every character within both words. Words that have the same stem are also considered as literal matching. Partial matches are also accepted if the word is long (at least 6 characters) and at least 70% of the characters were matched.

Soundex matching (Christian 1998) is an algorithm for finding a match between words by ignoring vowels and assigning characters that have similar pronunciation to the same soundex symbol (e.g. *b* and *p* are pronounced similarly, and are thus grouped together). Soundex matching helps to address any potential misspellings.

Detection of Reading Strategy Skills

R-SAT assessments have been designed to detect the use of individual comprehension strategies. However, readers invariably use multiple strategies when thinking aloud. Therefore, protocols could be considered in two ways: *individual strategy* use and *combinations of strategies* use. As such, we explored the accuracy of word matching and LSA to detect (i) individual strategies (e.g., use of an elaboration strategy) and (ii) a combination of strategies (use of both an elaboration and distal bridging strategy). Although detecting combinations of strategies is considerably more complex for NLP, it enables the system to track students' performance and adaptively respond (Rus et al. 2009). Consequently, this work is important because it could lead to improved feedback for students and it could also provide insight into the complex interplay between strategy use and comprehension.

Study

A total of 158 participants read six texts (2 science, 2 history, and 2 narrative). From these readings, 2,357 typed think-aloud protocols were collected and coded by trained human judges (who were graduate students majoring in psychology). For each protocol, the judges evaluated the presence or absence of each strategy on a 3-point scale:

0 - strategy is not present
1 - there is a noun phrase that reflects an information source associated with the strategy
2 - there is at least one verb clause that reflects the strategy.

Table 1 shows number and percent of protocols that were labeled 0, 1, or 2 for each reading strategy.

Table 1. Human Scoring on Dataset

Strategy	0	1	2
Paraphrase	445 (19%)	822 (35%)	1090 (46%)
Local Bridge	1049 (45%)	473 (20%)	835 (35%)
Distal Bridge	472 (20%)	354 (15%)	1531 (65%)
Elaboration	1039 (44%)	104 (4%)	1214 (52%)

For each protocol, the following variables were obtained:

- Protocol length, in number of words [cnt]
- Word matching values: protocol compared to …

- Current Sentence (Paraphrase) [para_words]
- Prior Sentence (Local bridging) [local_words]
- All Distal sentences (Global bridging) [distal_words]
- New Words - not in current sentence or prior discourse (Elaboration) [not_mentioned_words]

- LSA cosine values: protocol compared to …
  - Current Sentence (Paraphrase) [para_lsa]
  - Prior Sentence (Local bridging) [local_lsa]
  - All Distal sentences (Global bridging) [distal_lsa]
  - Subsequent sentences (Elaboration) [elab_lsa]

Thus, a total of 9 variables were computed for the purpose of these analyses.

The data were randomly divided into 3 groups of approximately equal size: one group was used to create a model (training set) and the other 2 groups are used as a validation of the model (test sets). Each group contains almost equal number of protocols labeling 2 (on a 3-point scale) in each reading strategy.

Individual Strategies: Effect on Size of Analysis

The first step toward detecting comprehension processes is to identify individual strategies that were used in each think-aloud protocol. As described above, the human coding scheme was designed on a 3-point scale. The three distinct levels provide variability in grain-size for the NLP algorithms. This could have significant implications for the accuracy of any NLP models. As an alternative to having three distinct levels, a dichotomous categorization scheme was also employed that collapsed across the original set of categories. The dichotomous scale consisted of collapsing the first two coding categories, so that the new coding consisted of a "0" if the strategy was not explicitly present and a "1" if the strategy was fully present. A protocol will be coded as fully present for a given strategy when it contains at least one verb clause that reflects the strategy. It is worth noting that a collapse of the latter two coding categories was evaluated, but the results were omitted in this paper.

A step-wise discriminant function analysis was used to compare the performance between several NLP models and the human ratings. Each model was assessed on both the full and collapsed rating schemes (3 or 2 categories, respectively)

- Word matching model: only word matching variables with protocol length (para_words, local_words, distal_words, not_mentioned_words, and cnt),
- LSA-based model: only LSA-cosine variables (para_lsa, local_LSA, distal_LSA, and elab_LSA), and
- Mixed model: combined all 9 variables from both the word-matching model and LSA-based model.

Various models were created using the test set. In Table 2 for each model, a list of significant variables in the model along with percent correctly classified on the train set, and F value at $p < 0.001$. A protocol is considered correctly classified when an NLP model produces the same results as human codings. For example, if human coding said a paraphrase strategy is present (i.e. value 1), an NLP model should give a value of 1 for a correctly classification.

Table 2. Model constructions and values from training set

Model	S	Significant Variables	% correct	F, $p < 0.001$
WM 3pt	P	para_words, local_words	59.8%	F(2,779) = 82.86
	L	cnt, para_words, local_words	66.3%	F(3, 778) = 78.51
	D	All, but cnt	60.7%	*F*(4, 776) = 39.33
	E	All, but elab_words	45.6%	F(4, 777) = 12.41
WM 2pt	P	para_words, local_words	71.6%	F(2, 779) = 135.98
	L	All, but elab_words	78.4%	F(4, 777) = 99.08
	D	local_words, distal_words, elab_words	73.9%	F(3, 778) = 91.86
	E	All, but elab_words	65.4%	F(4,776) = 25.42
LSA 3pt	P	All, but *distal_lsa*	46.4%	F(3, 778) = 29.50
	L	All, but para_lsa	56.5%	*F*(3, 778) = 45.28
	D	All, but para_lsa	59.0%	*F*(3, 778) = 44.42
	E	local_lsa	52.2%	*F*(1, 780) = 14.71
LSA 2pt	P	para_lsa, local_lsa	64.5%	*F*(2, 779) = 50.60
	L	local_lsa, elab_lsa	69.5%	*F*(2, 779) = 79.49
	D	distal_lsa, elab_lsa	68.5%	*F*(2, 779) = 100.11
	E	local_lsa, distal_lsa	55.3%	*F*(2, 779) = 15.83
Mixed 3pt	P	para_words, para_lsa, local_lsa, later_lsa	62.5%	F(4, 777) = 42.82
	L	cnt, local_words, elab_words, local_lsa, distal_lsa	66.2%	F(5, 775) = 52.36
	D	para_words, local_words, distal_words, elab_words, distal_lsa	62.9%	F(5, 775) = 36.76
	E	cnt, para_words, local_words, distal_wrods	44.8%	F(4, 776) = 13.05
Mixed 2pt	P	para_words, para_lsa, local_lsa	74.1%	F(3, 778)=103.30
	L	cnt, para_words, local_words, local_lsa, distal_lsa	78.4%	F(5, 775)=77.94
	D	local_words, distal_words, elab_words, distal_lsa	73.0%	F(4, 776) = 75.98
	E	cnt, para_words, local_words, distal_words, local_lsa, later_lsa	65.3%	F(6, 774) = 18.04

The combined model results had stronger performance. The human codings were also compared to the combined model performance (LSA and word matching together). Percent Agreement (%), and Unweighted Kappa(K) for this analysis are shown in Table 3.

These results suggest that reading strategies can be adequately detected within student protocols using NLP models. Unfortunately, all of the models had slightly lower performance when attempting to identify the presence of elaboration. One reason the elaboration strategy may be difficult to detect is that by definition, elaborations involve going beyond the textual context, and include personal relations between concepts. Thus, a successful elaboration can exist within an extremely open option space where students can include a practically infinite number of relevant responses.

The results between the two ratings schemes indicate that a more fine-grained, subtle approach may be the best method for identifying specific strategy use. Model performance dropped when transitioning from the three item scheme down to the two item dichotomous scheme. This drop was most evident for the text-based strategies of paraphrasing and bridging (distal bridging in particular). We believe that this difference may be the case because a score of 1 (strategy represented by a noun phrase) could be psychologically meaningful and reflect the process of anaphora resolution.

This trend is not reflected within the elaboration category; however that is somewhat expected given the personalized, non-text-based nature of that strategy. The model differences between ratings schemes have profound implications for future studies that compare NLP algorithms to human ratings. Namely, the grain-size of any human ratings scheme can have a significant effect on the accuracy outcomes for NLP models.

Table 3. Results of mixed models (LSA + WM) to predict reading strategies.

Strategy		3-point scale			2-point scale		
		Full set	Training set	Test set	Full set	Training set	Test set
P	%	0.593	0.616	0.582	0.515	0.503	0.520
	K	0.382	0.413	0.367	0.426	0.464	0.406
L	%	0.665	0.656	0.670	0.576	0.560	0.584
	K	0.477	0.466	0.483	0.551	0.529	0.561
D	%	0.622	0.612	0.626	0.480	0.481	0.479
	K	0.387	0.364	0.398	0.273	0.266	0.277
E	%	0.485	0.490	0.482	0.493	0.488	0.496
	K	0.195	0.205	0.189	0.300	0.310	0.294

Combined Reading Strategies

Although there is merit in detecting individual strategies, students rarely use a single strategy in isolation. It is more common for students to use a combination of strategies to help construct meaning (e.g., Trabasso and Magliano, 1996). Therefore, R-SAT needs to be able to account for the presence of multiple strategies within a single protocol. Hence, an analysis was performed that investigated combinations of reading strategies.

The human ratings were recoded to indicate all strategies present within each protocol. The new coding consisted of a four-character combination, XXXX, which represents each strategy present. Each letter position corresponds to one of the specified reading strategies (1-Paraphrasing, 2-Local bridging, 3-Distal bridging, and 4-Elaboration). If a strategy was coded as a 2 (in 3-point scale), then "X" will be replaced with the letter representing that strategy. For example, if a protocol included both paraphrasing and distal bridging, then that protocol would have the code "PXDX". With 4 reading strategies, a total of 16 possible categories were created:

1	2	3	4		15	16
XXXX,	XXXE,	XXDX,	XLXX,	…. ,	PLDX,	PLDE

A step-wise discriminant function analysis using the combined LSA and word matching algorithm to predict the sixteen human categories produced a significant model, $F(5, 776) = 16.346$, $p < 0001$. Table 4 shows the Percent Agreement (%) and Unweighted Kappa between the predicted category and human-coded category.

Table 4: Results of mixed model to predict combined strategies.

		2-point scale, combined strategies		
		Full set	Training set	Test set
Exact	C	0.605	0.646	0.583
	%	0.277	0.300	0.265
	K	0.220	0.246	0.208
+/-1	%	0.420	0.446	0.407

Using natural language to predict discrete category membership is typically difficult to achieve. In this particular analysis we were attempting to predict membership within sixteen distinct categories. Despite finding a relatively high kappa score for such a large number of categories, we conducted a more lenient follow-up analysis that included fuzzy category membership. In this additional analysis we accepted classifications as correct if they appeared within neighboring categories (an error of +/- 1). For example, if humans rated a protocol as "PXDX", then the lenient coding would accept "PLXX", "PXDX", or "PXXE" as correct. This lenient analysis indicates a significant increase in model performance (kappa increases from .220 to .420).

Detection of Reading Comprehension

The previous analyses demonstrate an ability to detect the presence of specific reading comprehension strategies. Taking this a step further, we wanted to investigate if the presence of these reading strategies is related to student performance.

The Gates-MacGinitie Reading Test (GMRT) is a standardized test used to measure reading comprehension. In addition to the verbal protocols each student in this study completed the GMRT. The aforementioned NLP models were used in an attempt to predict reading comprehension scores. The models used in this analysis include the following: word matching only, LSA only, words and LSA combined, 3-point scale, 2-point scale, and combined 16 strategies.

Score Aggregation

GMRT scores are computed for each student. Therefore aggregate scores for each variable were calculated for each student.

Individual Strategies: An average value was calculated for each of the nine variables within the word matching and LSA models. The human codings were also averaged to create an aggregate score for each student.

Combined Strategies: Student scores for strategy combinations consisted of a total frequency count for each of the sixteen categories.

Models for predicting GMRT score

Nine models were used to predict comprehension score. These models include the words only algorithm, LSA only algorithm, mixed NLP algorithm, original human coding (3-point scale), collapsed human coding, combined human coding (16 strategy combinations), NLP predicted original human coding, NLP predicted collapsed human coding, and NLP predicted combined human coding.

Word Matching model (M1-WM) The word matching variables were used in an attempt to predict the GMRT scores. A step-wise regression analysis indicated only one significant variable, *para_words*, $F(1, 52) = 15.218$, $p < 0.001$. When all word matching variables were used in the backward regression analysis, $F(5, 48) = 3.784$, $p < 0.01$.

Table 5. Pearson correlations between individual NLP measures

	1	2	3	4	5	6	7	8	9	10
1. cnt	1	0.61	0.80	0.90	0.86	0.31	0.60	0.81	0.74	0.27
2. para_words		1	0.59	0.62	0.35	0.77	0.43	0.60	0.61	0.35
3. local_words			1	0.78	0.55	0.36	0.79	0.74	0.66	0.30
4. distal_words				1	0.63	0.29	0.56	0.89	0.67	0.22
5. elab_words					1	0.12	0.41	0.58	0.62	0.23
6. para_lsa						1	0.37	0.33	0.44	0.36
7. local_lsa							1	0.53	0.54	0.34
8. distal_lsa								1	0.64	0.34
9. later_lsa									1	0.39
10. GMRT										1

LSA-based model (M2-LSA) The LSA variables were included within a step-wise regression analysis to predict reading comprehension score. This analysis found two significant variables, *para_lsa* and *later_lsa*, $F(2, 51) = 10.889$, $p < 0.001$. When all LSA variables were used in the backward regression analysis, $F(4, 49) = 6.903$, $p < 0.001$.

Mixed model (M3-Mixed) The combination of word matching and LSA variables were used to predict comprehension scores. A step-wise regression discovered only a single significant variable, *para_words*, $F(1, 52) = 15.218$, $p < 0.001$. When all 9 variables (both word matching variables and LSA variables) were used in the backward regression analysis, $F(9, 44) = 3.130$, $p < 0.01$.

Human 3-point scale model (M4-Human 3-point) The original human ratings (on a 3-point scale) were used in model four to predict reading comprehension scores. This model does not include any NLP variables. A step-wise regression analysis yielded only one significant variable, *paraphrase skill*, $F(1, 52) = 25.085$, $p < 0.001$. When all 4 strategy skills (paraphrase, local bridge, distal bridge, and elaboration) were used in the backward regression analysis, $F(3, 49) = 7.554$, $p < 0.001$.

Human 2-point scale model (M5-Human 2-point) Model five used the collapsed (dichotomous scale) humans ratings to predict GMRT scores. This model also does not include any NLP variables. A step-wise regression analysis resulted in only one significant variable, *paraphrase skill*, $F(1, 52) = 26.848$, $p < 0.001$. When all 4 strategy skills were used in the backward regression analysis, $F(4, 49) = 9.077$, $p < 0.001$.

Human Combined Strategy model (M6-Human Combined Strategy) The complete set of sixteen categories were used to predict the reading comprehension scores. Again, this model does not include any NLP variables. The step-wise regression analysis is uncovered only one significant variable, *only distal bridging strategy (category 3)*, $F(1, 52) = 19.270$, $p < 0.001$. When all 16 strategy skills were used in the backward regression analysis, *only elaboration strategy (category 2)* was excluded from the model, $F(15, 38) = 2.659$, $p < 0.01$.

Three NLP models were created that used the predicted human scores. In models M7-M9, the formula created for model M4-M6 were used, but the values of strategy skills (3-point or 2-point) were from the predicted strategies instead of human. For example, M7 will use the formula created by M4. However, instead of testing it using

human-coding strategy skill, we will use the predicted value instead. These values were calculated in *Individual Strategies* section, described above. Similar concept applied to M8 that uses M5's formula and M9 that uses M6's formula.

The correlations between the GMRT scores and the models are shown in Table 6.

Table 6. Correlation between predicted score and gates

Model	Full set	Training set	Test set
M1-WM	0.364	0.532	0.281
M2-LSA	0.393	0.600	0.308
M3-Mixed	0.433	0.625	0.351
M4-Human 3-point	0.530	0.618	0.484
M5-Human 2-point	0.533	0.652	0.469
M6-Human Combined Strategy	0.512	0.716	0.416
M7-Predicted Skills, 3- point	0.375	0.480	0.318
M8-Predicted Skills, 2-point	0.433	0.501	0.397
M9-Predicted Combined Strategy	0.204	0.463	0.056

These results demonstrate that reading comprehension can be moderately predicted from the reading strategy measures. This suggests that if we can accurately identify the presence of specific strategies, then we can more adequately estimate a student's reading ability. It is confirmed that good predicted strategy skills (either 3-point or 2-point) will result in a better prediction in reading comprehension.

Discussion

This study addressed the use of various NLP measures for identifying specific reading comprehension strategies. These strategies can be used in isolation or in combination. The current analysis indicates that we can be moderately confident in detecting individual strategies. Extending beyond single strategies to try and correctly identify any possible combination of strategies is clearly a difficult task.

The current study discovered potential limitations for future NLP analyses involving human ratings. Analyses utilizing 2 variations of a rating scheme indicated that the fine-grained scheme provided improved model performance. Specifically, the model's ability to detect the presence of distal bridging declined markedly when two of the categories were collapsed. This performance dip may be explained by the specific nature of bridging (and any associated anaphora resolution). This finding indicates that model evaluations are highly contingent upon the structure of human comparisons.

Additional analyses were conducted to investigate the relation between student reading comprehension scores and reading strategy usage. These results found that the presence of reading strategies can account for a significant amount of the variance for reading comprehension scores. Further analyses indicate that R-SATs ability to predict strategy use also allows it to predict a portion of students' reading comprehension scores.

The primary goal of this work was to assess R-SAT's ability to detect strategy use. The current results indicate that a mixed model of both word matching and LSA is the most effective method for strategy classification. Continued work is planned that will investigate potential strategy differences between text genre as well as strategy use patterns that may emerge over time.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305G040055) to Northern Illinois University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Christian, P. (1998). Soundex - can it be improved? *Computers in Genealogy, 6*(5).

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., and Boonthum, C. (2007). Assessing the format of the presentation of text in developing a reading strategy assessment tool (R-SAT). *Behavior Research Methods*, *39*(2), 34-44.

Landauer, T.K. (2007). LSA as a Theory of Meaning. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–35). Mahwah, NJ: Erlbaum.

Landauer, T., McNamara, D.S., Dennis, S., and Kintsch, W. *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Magliano, J.P. and Millis, K.K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction. 21*, 251-283.

Magliano, J.P., Millis, K.K., The R-SAT Development Team, Levinstein, I.B., and Boonthum, C. (in press). Assessing Comprehension During Reading with the Reading Strategy Assessment Tool (R-SAT). *Metacognition and Learning*.

McNamara, D. S., Boonthum, C., Levinstein, I. B., and Millis, K. K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis,* (pp. 224-241) Mahwah, NJ: Erlbaum.

Rus, V., McCarthy, P.M., McNamara, D.S., and Graesser, A.C. (2009). Natural language understanding and assessment. In J.R. Rabuñal, J. Dorado, A. Pazos (Eds.). *Encyclopedia of Artificial Intelligence*. Hershey, PA: Idea Group, Inc.