Are You Reading My Mind? Modeling Students' Reading Comprehension Skills with Natural Language Processing Techniques

Laura K. Allen
Arizona State University
PO Box 872111
Tempe, AZ, 85287
01+404-414-5200
LauraKAllen@asu.edu

Erica L. Snow
Arizona State University
PO Box 872111
Tempe, AZ, 85287
01+404-414-5200
Erica.L.Snow@asu.edu

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ, 85287
01+404-414-5200
Danielle.McNamara@asu.edu

ABSTRACT
This study builds upon previous work aimed at developing a student model of reading comprehension ability within the intelligent tutoring system, iSTART. Currently, the system evaluates students' self-explanation performance using a local, sentence-level algorithm and does not adapt content based on reading ability. The current study leverages natural language processing tools to build models of students' comprehension ability from the linguistic properties of their self-explanations. Students (n = 126) interacted with iSTART across eight training sessions where they self-explained target sentences from complex science texts. Coh-Metrix was then used to calculate the linguistic properties of their aggregated self-explanations. The results of this study indicated that the linguistic indices were predictive of students' reading comprehension ability, over and above the current system algorithms. These results suggest that natural language processing techniques can inform stealth assessments and ultimately improve student models within intelligent tutoring systems.

Categories and Subject Descriptors
K.3.1 **[Computer Uses in Education]** *Computer-assisted instruction (CAI)*; I.2.7 **[Natural Language Processing]** *Text analysis, discourse*; J.5 [**Computer Applications: Arts and Humanities]**: Linguistics

General Terms
Algorithms, Measurement, Performance, Languages, Theory

Keywords
Intelligent Tutoring Systems, Natural Language Processing, stealth assessment, corpus linguistics, reading comprehension

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA
Copyright 2015 ACM 978-1-4503-3417-4/15/03…$15.00
http://dx.doi.org/10.1145/2723576.2723617

1. INTRODUCTION
Intelligent tutoring systems (ITSs) are designed to provide adaptive instruction and feedback to students based on their individual skills, levels of prior knowledge, and attitudes [1]. Typically, the focus of these systems lies in the adaptability of student feedback through automated assessments of student performance. However, many ITSs (particularly those focused on instruction in ill-defined domains) lack adaptive *instruction* and higher-level feedback, as it can be difficult to determine students' ability levels without subjecting them to extensive pretest measures.

One solution to this assessment problem lies in the use of natural language processing (NLP) techniques to assess the properties of students' natural language input [2]. Increasingly large numbers of ITS developers have begun to incorporate NLP within their systems [3-5], as it affords ITSs the opportunity to quickly and accurately evaluate the quality and content of students' responses [6-8]. Further, these evaluations allow systems to track students' skill and knowledge development, and to integrate this information with system data to develop models of student learners [9].

NLP components have been successfully integrated into ITSs for the purposes of increasing interactivity and learning outcomes. For instance, the AutoTutor system simulates the dialogue moves of human tutors to help students learn Newtonian physics, computer literacy, and critical thinking topics [3,10-12]. In AutoTutor, animated agents engage students in a dialogue to help students generate explanations of the key concepts presented by the system. Animated agents provide feedback on natural language responses and then guide students to develop their answers more deeply [3]. ITSs that similarly utilize NLP to interact with users are Why2-Atlas, which tutors students on physics problems [13-14] and iSTART, which tutors students on reading comprehension strategies [15].

Prior research suggests that interactions with NLP-based tutoring systems lead to significant learning gains. Compared to non-interactive learning tasks, such as reading textbooks or listening to lectures, these NLP-based ITSs provide greater performance gains across time [5,16]. Graesser, Jackson and colleagues (2003), for instance, investigated the benefits of the AutoTutor system in comparison to non-interactive tasks [17]. Three conditions were compared in this study: tutoring from AutoTutor, reading a

physics textbook, and a control condition with no educational content. The results indicated that students in the AutoTutor condition outperformed students in the other two conditions. Further, results revealed that the two non-AutoTutor conditions produced equivalent learning gains. This suggests that the non-interactive reading task was equivalent to the absence of any sort of learning task.

Despite these positive findings, NLP-based systems still have plenty of room for improvement. One shared difficulty lies in the ability of these systems to represent students' knowledge and overall ability levels. ITSs typically assess students' abilities by evaluating performance on individual items, such as math problems or the quality of students' explanations of a particular concept. This is a beneficial form of assessment, as it allows systems to provide item-level feedback, as well as to adapt content. One problem, however, is that these assessments provide little indication of students' prior abilities at a more global level. In order to provide the most beneficial instruction and feedback, ITSs should assess students' ability levels on numerous dimensions. The current paper extends previous research that aimed to model students' reading comprehension abilities in iSTART [18]. Here, we utilize an automated text analysis tool to model students' reading comprehension skills through an analysis of the linguistic properties of their self-explanations.

1.1 Stealth Assessments and Student Models

ITSs are typically developed to rely on frequent assessments of users' performance, affect, and skills as they progress through the system. Despite the importance of these measures, however, researchers and educators frequently debate the optimal frequency and timing of these assessments. On one hand, it has been suggested that system designers should try to avoid the persistent questioning of users, as these assessments can easily disrupt learners' "flow" during training tasks [19]. On the other hand, it is difficult to ignore the fact that these assessments can drastically increase the effectiveness of ITSs, as they can lead to enhanced personalization and adaptability of the system's feedback and instruction. As a result of this assessment problem, researchers and developers have been tasked with developing novel methods for collecting information about student users that do not consistently disrupt the learning task [19,20].

One method that has been proposed in response to this issue is the development of "stealth" assessments [19,20]. Stealth assessments are "invisible" measures that are designed to collect student information (e.g., their level of engagement, current affect, cognitive skills, etc.) without subjecting learners to explicit tests. These covert measures are typically embedded within specific learning tasks and are, therefore, not able to be detected by the students [21]. Within ITSs, there is a wealth of information that can be easily collected and used to inform these stealth assessments. In-system log data (e.g., students' clicks and system choices), for example, has been used to discretely measure students' levels of engagement during various learning tasks [22]. Such data can include a wide variety of information, such as the choices students while interacting with the system or the characteristics of the natural language they generate within the system.

An important and beneficial characteristic of stealth assessments is that they are not constrained like more traditional self-report measures that rely on students' perceptions or memories of a particular learning task. Indeed, one concern that commonly arises from the use of traditional measures is that they may not fully or adequately capture their target construct [23]. Students may claim to have felt a certain emotion or learned a specific skill while engaged in the system; however, it has been well documented that students' perceptions of their performance and their actual performance are often misaligned [23, 24]. Stealth assessments can "side step" this issue by specifically measuring the target behaviors as they occur *naturally*. Thus, because these measures capture actual behaviors during the learning process, they do not rely on the accuracy of students' depictions of their own learning.

Stealth assessments can take many forms (i.e., log data, accuracy, mouse movements, etc.), and can be used to measure a wide variety of constructs (i.e., behaviors, motivations, competency, etc.). Once developed, the greatest benefit of these assessments is that they can be used to inform student models. ITS developers frequently embed student models within their systems as a way to provide individualized instruction and feedback to students [25]. This personalization is driven by a continuously updated model that represents students' knowledge and performance within the system.

Once the system is able to reliably assess an individual student's particular strengths and weaknesses, it can adapt in ways that will increase the efficiency of the training. As an example, consider the ITS iSTART, which provides students with instruction and training on self-explanation reading strategies. If a student within iSTART can be identified as having weak vocabulary skills, the system might initially assign texts with more familiar, concrete words compared to texts potentially assigned to student with stronger vocabulary knowledge. If an initially weaker student begins to demonstrate a stronger degree of vocabulary knowledge, however, the system can respond to this development and increase the difficulty of the words in the assigned texts.

Importantly, stealth assessments provide a number of benefits over more traditional assessments of students' performance. First, stealth assessments allow students to be continuously assessed without disrupting the learning process. This information then increases the opportunities for the system to adapt to students based on their specific pedagogical needs [26]. Additionally, stealth assessments are based on students' *natural* behaviors while engaged with learning tasks. Thus, they do not rely on faulty, post-hoc measurements or self-reports of performance; rather they use information that was generated during training to serve as a proxy for students' performance or affect. Overall, by covertly collecting a wealth of information about behaviors and skills, ITSs can provide a more direct and individualized experience to each learner without interrupting their learning paths.

1.2 iSTART

The Interactive Strategy Training for Active Reading and Thinking (iSTART) tutor is an ITS developed to train high school and college students on the use of reading comprehension strategies [15]. In particular, iSTART focuses on the instruction of self-explanation strategies, which have previously been shown to be beneficial for various high-level skills including problem solving, generating inferences, and deep comprehension of texts [27-28]. iSTART is divided into three training modules: introduction, demonstration, and practice. Within the introduction module, students are given a brief overview of the self-explanation reading strategies and provided examples of these strategies. In the demonstration module, students watch as two pedagogical agents demonstrate how to apply the self-explanation strategies to complex science texts. Finally, in the practice

module, students are provided with the opportunity to apply the strategies they have learned to new texts.

The development of the iSTART tutor was based on previous research that had been conducted using a human-based self-explanation training intervention called SERT (Self-Explanation Reading Training) [28]. The purpose of this intervention was to provide strategy instruction and training to students who struggled to comprehend complex texts. Previous research suggests that both SERT and iSTART are effective at improving students' ability to comprehend difficult science texts [29].

Training in iSTART is separated into three modules (i.e., introduction, demonstration, and practice), which map onto three pedagogical principles – modeling, scaffolding, and fading instruction. In the introduction module, students are introduced to self-explanation strategies (comprehension monitoring, prediction, paraphrasing, elaboration, and bridging). These strategies are explained to students via a vicarious conversation between a teacher agent and two student agents. Throughout this exchange between the teacher and student agents, the self-explanations are described and examples of these strategies are provided.

After interacting with the introduction module, students move on to the demonstration training module (see Figure 1 for a screenshot of this module). In this module, the student is able to see the various reading strategies applied to an example text by two animated agents (one teacher and one student). The student agent first applies the iSTART reading strategies to target sentences within the text. Then, the tutor agent provides feedback on the quality of the example self-explanation. The user is prompted to identify the specific strategy that was used in the example. The dialogue between the two agents serves to preview the interactions that will take place between the user and the iSTART system in the practice module.

In the practice module, students are given the opportunity to apply the strategies to new texts. During the practice module, direct instruction is faded out and students are required to provide more information and interactions. Throughout practice, the teacher agent from demonstration serves as a 'self-explanation coach' by providing students with feedback on their self-explanations and prompting the use of other strategies.

The practice module in iSTART is divided into two separate sections. The first practice module (initial practice) is housed within the initial two-hour training portion of iSTART. During initial practice, students apply strategies and receive feedback on their self-explanations using two different texts. The second practice module (i.e., extended practice) is designed to provide a long-term environment for strategy practice that can last for weeks or months. In this environment, students are asked to self-explain target sentences within a variety of science texts. In the extended practice module, teachers have the opportunity to input and assign new texts for their students to practice. Thus, the iSTART assessment algorithm must be flexible so that it can evaluate self-explanations that have been produced for any text.

1.2.1. iSTART Evaluation Algorithm

iSTART utilizes a localized (i.e., sentence-level) evaluation algorithm that assesses the quality of individual self-explanations and allows the system to provide relevant feedback [4].

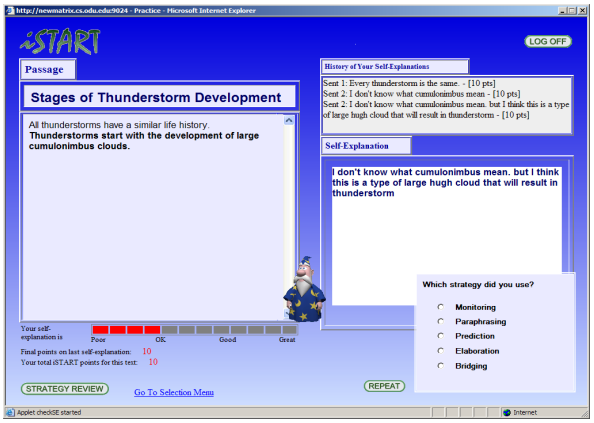

Figure 1. Screenshot of iSTART demonstration module

Currently, the algorithm assesses self-explanations based on information from a relatively small window. This window includes the specified target sentence from the text that students self-explain and the surrounding sentences from the target passage. This algorithm assesses self-explanation quality using multiple word-based indices and latent semantic analysis (LSA) [30]. Lower-level information about the self-explanations is provided by the word-based indices, such as length and overlap in the content words. These measures help the system detect self-explanations that are too short, too similar to the topic sentence, or irrelevant. LSA is then combined with these measures to provide a more holistic index of quality. This index provides information about the degree to which self-explanations include information from earlier in the text and from outside the text (i.e., prior world knowledge).

iSTART utilizes the information from these word-based and LSA-based indices to provide self-explanations a score from 0 to 3. A score of "0" indicates that a self-explanation is too short to accurately assess or that it contains irrelevant information. A score of "1" is assigned to self-explanations that are primarily related to the target sentence only. A score of "2" implies that a self-explanation integrates information from the text beyond the target sentence. Finally, a score of "3" is associated with self-explanations that incorporate outside information at the global level; thus, this self-explanation can include information that was not included in the text, or a self-explanation that focuses on overall themes that persist throughout the text. Previous research suggests that the iSTART algorithm can perform as accurately as humans and provides a general estimate for the amount of cognitive processing that was required to generate the self-explanation [31].

1.3 iSTART-ME

Research suggests that the initial and extended practice modules in iSTART lead to increases in students' ability to self-explain and comprehend complex science texts [32-33]. However, these studies also indicate that students frequently disengage from training, as the modules can easily become repetitive over time [34-35].

Figure 2. Screenshot of iSTART-ME selection menu

To address student disengagement, educational games and game-based features were incorporated within the system to develop iSTART-ME (iSTART-Motivationally Enhanced). Within iSTART-ME, the three main modules of iSTART (i.e., introduction, demonstration, and practice) remain relatively unchanged. However, the extended practice module contains a selection menu that allows students to interact with both educational games and game-based features (see Figure 2 for a screenshot of the game-based selection menu).

Throughout extended practice, students can earn iSTART points by interacting with three different types of generative practice: Coached Practice, Showdown, and Map Conquest. Coached Practice is the non-game-based method of self-explanation practice from the original version of iSTART. The two other practice methods, Showdown and Map Conquest, are game-based forms of generative practice that use the same NLP algorithm from Coached Practice. As students interact with any of these practice environments, they earn points within the system and, advance to higher achievement levels.

The iSTART-ME interface contains additional game-based features; however, these features are not relevant to the current study (for a more detailed description of all these game-based features, see Jackson, Dempsey, & McNamara, 2010) [36]. Overall, the iSTART-ME system has been shown to increase student enjoyment and learning throughout strategy training [35].

1.4 Current Study

The purpose of the current study is to investigate the degree to which the linguistic properties of students' self-explanations can inform stealth assessments of their reading comprehension abilities. Ideally, these assessments will serve to inform student models in the iSTART system and contribute to its adaptability in the form of more sophisticated scoring algorithms, feedback, and adaptive instruction. To this end, students' self-explanations from the iSTART and iSTART-ME systems were collected. These individual, sentence-level self-explanations were then aggregated across each of the texts that were read and analyzed using Coh-Metrix. Coh-Metrix is an automated text analysis tool that provides linguistic indices related to the lexical sophistication, syntactic complexity, and cohesion of texts. We used this tool in the current study so that we could examine relationships between students' reading comprehension skills and the linguistic characteristics of their natural language input within iSTART. We hypothesized that these linguistic indices would be positively related to reading comprehension ability and, importantly, provide added predictive power over and above the current system information (i.e., the local NLP algorithm).

2. METHODS

2.1 Participants

Participants were 126 high-school students from a mid-south urban environment. Of these students, 65 interacted with the iSTART-original system and 61 interacted with the iSTART-ME system. Because students in both conditions completed the same cognitive tasks and were assessed via the same algorithm, the current analyses are collapsed across both conditions. All students were monetarily compensated for their participation in this experiment.

2.2 Procedure

The current study was an 11-session experiment that consisted of a pretest, 8 training sessions, a posttest, and a delayed retention test. During the first session, students completed the pretest, which included measures of their reading comprehension skills, self-explanation ability, and affective states. Training occurred during the following eight sessions in which students engaged with the iSTART-original or iSTART-ME system. During session 10, students completed a posttest, which comprised measures similar to the pretest.

2.3 Reading Comprehension Assessment

Students' reading comprehension skills were measured using the Gates-MacGinitie (4th ed.) reading skill test (form S) level 10/12 [37]. This assessment contained 48 multiple-choice questions that measured students' ability to comprehend shallow and deep level information across 11 short text passages.

2.4 Text Analyses

The linguistic features of students' aggregated self-explanations were calculated using Coh-Metrix. These features are discussed in greater detail below.

2.4.1 Coh-Metrix

To assess the linguistic properties of students' aggregated self-explanations, we utilized Coh-Metrix. Coh-Metrix is an automated text analysis tool that computes linguistic indices for both the lower- and higher-level aspects of texts [38]. These indices range from basic text properties to lexical, syntactic, and cohesive measures.

Coh-Metrix calculates a number of basic linguistic indices, which provide simple counts of features in a given text. This category includes descriptive indices, such as the total number of words, parts of speech (e.g., nouns, verbs, etc.), and paragraphs in a given text. In addition, Coh-Metrix calculates the average length of words (average number of letters and syllables), sentences (average number of words per sentence), and paragraphs (average number of sentences per paragraph) within that text.

The lexical indices calculated by Coh-Metrix describe the characteristics of the words that are found in a given text. Examples of these indices include lexical diversity (i.e., the degree to which the text contains unique words, rather than repetitive language), and the age of acquisition of the words (i.e., the age at which the words used in the text are typically acquired by children).

Syntactic measures describe the complexity of the sentence constructions found within a text. Examples of these indices include the number of modifiers per noun phrase and the

incidence of sentences that are agentless and constructed in the passive tense.

Cohesion measures provide information about the type of connections that are made between ideas within a text; some relevant measures include: incidence of connectives, minimal edit distance (MED; indicates the structural similarity of the sentences in a text), and content word overlap (for adjacent sentences and all sentences). Finally, Latent Semantic Analysis (LSA) is used to provide information about the semantic similarity of texts. LSA is a statistical and mathematical representation of word and text meaning that uses a technique called singular value decomposition to reduce a large corpus of texts into 300-500 dimensions [30]. The resulting dimensions are representative of the co-occurrence of words and phrases across a wide range of texts. Coh-Metrix provides numerous LSA measures for texts, such as the LSA paragraph-to-paragraph measure. This value indicates the semantic similarity of the concepts across the paragraphs of a given essay without relying on measures of morphological similarity.

2.5 Data Processing

For the purpose of calculating the linguistic properties of students' self-explanations, their individual, sentence-level self-explanations were aggregated for each of the texts that they read during their training. Therefore, each student had one "aggregated self-explanation" for each text that was read during iSTART training. As an illustration, for a target text with *p* paragraphs and *n* target sentences, the resulting aggregated self-explanation file would contain *p* paragraphs and *n* self-explanations corresponding to the relative position of the target sentence (see Figure 3 for visualization of this aggregation). This aggregation method has been discussed in more detail in previous research [18].

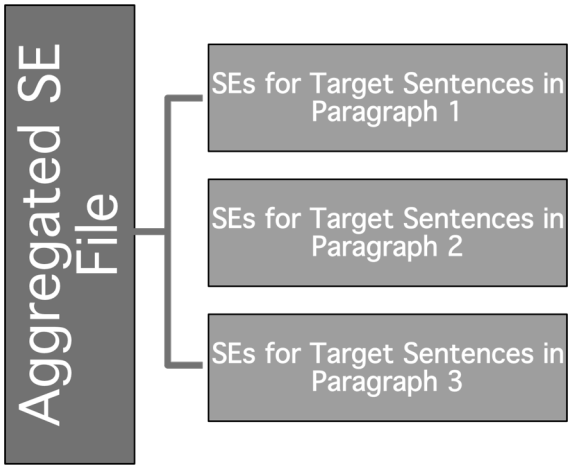

Figure 3. Visual of self-explanation aggregation

Linguistic indices were then calculated for each of these aggregated self-explanation files using Coh-Metrix. For each of the student users, this Coh-Metrix output was averaged across all of the texts in the system, which resulted in an average score for each of the linguistic measures. These average linguistic scores provide information about students' aggregated self-explanations at multiple linguistic levels (e.g., word-level, sentence-level, and passage-level information).

2.6 Statistical Analyses

Statistical analyses were conducted to investigate the role of linguistic text properties in assessing and modeling students' reading comprehension scores. Pearson correlations were first calculated between students' scores on a reading comprehension measure and the properties of their aggregated self-explanations. Indices that demonstrated a significant correlation with reading comprehension scores ($p < .05$) were retained in the analysis. Multicollinearity of these variables was then assessed among the indices ($r > .90$) – in the case that indices demonstrated multicollinearity, the index that correlated most strongly with reading comprehension scores was retained in the analysis. All remaining indices were finally checked to ensure that they were normally distributed. A stepwise regression analysis was then conducted to assess which of these linguistic properties were most predictive of reading comprehension abilities. For this regression analysis, a training and test set approach was used (67% for the training set and 33% for the test set) in order to validate the analyses and ensure that the results could be generalized to a new data set. Finally, a hierarchical regression analysis was conducted to determine whether the linguistic properties of the aggregated self-explanations accounted for additional variance over and above variance accounted for by performance as reflected by the iSTART algorithm score provided in the iSTART system [31].

3. RESULTS

3.1 Reading Comprehension Analysis

Pearson correlations were calculated between the Coh-Metrix linguistic indices and students' Gates-MacGinitie reading comprehension scores to examine the strength of the relationships among these variables. This correlation analysis revealed that there were 29 linguistic measures that demonstrated a significant relation with reading comprehension scores. However, 5 variables were removed due to strong multicollinearity with each other. The 24 remaining variables are included in Table 1 (see McNamara et al., 2014, for explanations of each variable) [38].

We calculated a stepwise regression analysis with these 24 Coh-Metrix indices as the predictors of students' reading comprehension scores for the 90 students in the training set. This regression yielded a significant model, $F(3, 86) = 17.624$, $p < .001$, $r = .617$, $R^2 = .381$. Three variables were significant predictors in the regression analysis and combined to account for 38% of the variance in students' comprehension scores: lexical diversity [$\beta = .38$, $t(3, 86) = 4.179$, $p < .001$], LSA paragraph-to-paragraph [$\beta = .32$, $t(3, 86) = 3.519$, $p = .001$], and average sentence length [$\beta = -.22$, $t(3, 86) = -2.532$, $p = .013$]. The regression model for the training set is presented in Table 2. The test set yielded $r = .519$, $R^2 = .269$, accounting for 27% of the variance in comprehension scores (see Table 2 for an overview of the regression analysis).

The results of this regression analysis indicate that the students with higher comprehension scores produced self-explanations that had more diverse word choices and shorter sentences. Despite using more diverse words, however, the skilled comprehenders' self-explanations contained a greater degree of semantic similarity than did those generated by less skilled comprehenders. Thus, these students may have been establishing relationships amongst the text concepts at a semantic level, rather than by simply repeating the same words and information.

Table 1. Correlations between Gates-MacGinitie reading comprehension scores and Coh-Metrix linguistic scores

Coh-Metrix variable	*r*	*p*
LSA paragraph-to-paragraph	.456	<.001
Lexical diversity	.452	<.001
Number of sentences	.445	<.001
LSA given/new	-.437	<.001
MED (all words)	.419	<.001
LSA (all sentences)	.362	<.001
Hypernymy (nouns)	.359	<.001
Number of words	.357	<.001
LSA adjacent sentences (standard deviation)	.362	<.001
Frequency of content words	.324	<.001
Temporal connectives	.314	<.001
Age of acquisition of words	.294	<.01
Average sentence length	-.261	<.01
Verb overlap (WordNet)	.258	<.01
Aspect repetition score	.247	<.01
Intentional ratio	-.246	<.01
Causal verbs	.243	<.01
Causal ratio	-.221	<.01
Second person pronouns	-.214	<.05
Third person pronouns	.199	<.05
Intentional events	.194	<.05
LSA paragraph-to-paragraph (standard deviation)	-.190	<.05
Verb overlap (LSA)	.187	<.05
Agentless passive constructions	.185	<.05

Table 2. Coh-Metrix regression analysis predicting Gates-MacGinitie reading comprehension scores

Entry	Variable added	R^2	ΔR^2
Entry 1	Lexical diversity	.254	.254
Entry 2	LSA paragraph-to-paragraph	.319	.081
Entry 3	Average sentence length	.381	.046

3.2 Comparison to Current Student Model

Our second analysis specifically tested the ability of the linguistic indices to predict students' Gates-MacGinitie comprehension scores over and above the scores provided by the current iSTART self-explanation algorithm. To address this question, we calculated a hierarchical multiple regression analysis to predict students' Gates-MacGinitie reading comprehension scores. Training scores (i.e., students' average self-explanation scores as assessed by the current iSTART algorithm) were entered into block 1 of the model, with the second block containing the three linguistic indices that were retained in the regression analysis above (lexical diversity, LSA paragraph-to-paragraph, and average sentence length).

Table 3. Hierarchical multiple regression analysis for linguistic variables predicting students' reading comprehension ability

Variable	*B*	SE *B*	β	ΔR^2
Model 1				.32**
Training Scores	12.40	1.61	.57**	
Model 2				.12**
Training Scores	8.66	1.78	.40**	
Lexical Diversity	.16	.05	.23**	
LSA paragraph-to-paragraph	12.99	7.911	.14	
Average sentence length	-.14	.05	-.19**	

** $p < .01$

The results of the hierarchical multiple regression analysis are presented in Table 3. Model 1 serves as a confirmation that the current iSTART algorithm is effective, as it significantly contributes to a model of students' reading comprehension ability and accounts for 33% of the variance. Model 2 provides a confirmation of our research hypothesis – namely, that the linguistic indices accounted for unique variance in reading comprehension scores over the current iSTART algorithm. Therefore, by analyzing the linguistic characteristics of students' self-explanations at multiple levels (i.e., at the word, sentence, and passage levels), we were able to improve the accuracy of the current iSTART student model.

4. DISCUSSION

Recent research suggests that intelligent tutoring systems (ITSs) are highly effective at producing learning gains among students – frequently performing as well as human tutors in the same domain [26]. One of the major difficulties that ITS developers still face, however, is the ability to process and respond to students' natural language input for the purpose of providing more adaptive learning experiences. Recently, developers of ITSs have begun to utilize NLP techniques to improve the adaptability of their systems [3-5]. While such NLP-based algorithms tend to be accurate and reliable at measuring performance on individual items, they have yet to inform more general models of student abilities.

In the current study, we used NLP techniques to develop stealth assessments of students' reading comprehension abilities. Specifically, an automated text analysis tool was used to analyze students' aggregated self-explanations of texts. This tool (Coh-Metrix) provided linguistic information about the self-explanations at multiple levels of the text. Importantly, the data calculated by this tool was able to significantly predict students' reading comprehension skills as assessed by the Gates-MacGinitie prior to training. Thus, by investigating students' self-explanations at the aggregate level, we are able to improve our current model of comprehension ability.

The Coh-Metrix correlation analysis revealed that a number of linguistic properties of students' self-explanations were related to their comprehension scores. Specifically, better readers' self-explanations were characterized by greater cohesion, shorter sentences, more connectives, greater lexical diversity, and more sophisticated vocabulary. Regression analyses revealed that lexical diversity, semantic cohesion (LSA paragraph-to-paragraph), and sentence length provided the most predictive

power in the model, accounting for 38% of the variance in students' reading comprehension scores. Hence, better readers tended to use a greater diversity of words and shorter sentences, possibly because they have more sophisticated vocabulary and were more likely to use punctuation within their explanations. Additionally, they maintained the topic more cohesively across their individual self-explanations of texts, as reflected by the LSA scores.

Importantly, the follow-up hierarchical regression analysis revealed that students' reading comprehension scores were positively related to both their training scores (current iSTART algorithm) and the linguistic indices of their aggregated self-explanations. The training scores accounted for a significant amount of variance in students' reading comprehension scores. This finding provides confirmation that the current iSTART algorithm is accurate and can contribute to a model of students' comprehension skills. Additionally, however, the results show that 12% of the *unique* variance in reading comprehension scores could be accounted for by three of the linguistic indices provided by Coh-Metrix. These findings suggest that there are passage-level linguistic properties of self-explanations that are important for modeling students' comprehension over and above quality information at only the sentence level. Thus, students' reading comprehension skills are better modeled when both the *quality* and the *properties* of their self-explanations are taken into consideration.

The results from the current study suggest that NLP measures can be used to provide stealth assessments of student abilities. When taken together, four of the Coh-Metrix variables accounted for almost half of students' reading comprehension ability. These findings are critical, as they indicate that students' skills can manifest in the ways in which they explain certain concepts within texts. Thus, linguistic analyses of self-explanations (and any user input) can provide information about students' processes as they engage in text comprehension. In this study, we only analyzed pretest reading comprehension skills; however, in the future, these methods could be applied to a model multiple student skills at different time points throughout training.

This study extends previous work, which showed that analyzing natural language input at both fine-grain and global sizes provided differential contributions to models of students' abilities [18]. The analyses presented here take this previous study a step further by using a multitude of fine-grained linguistic and semantic measures to develop an algorithm that can model approximately half of the variance in students' reading comprehension scores prior to iSTART training. Clearly, more work remains to extend these findings, namely in augmenting the feedback algorithm to consider students' prior abilities. Nonetheless, these analyses take a strong step towards increasing the adaptability and personalized instruction within the iSTART system. More broadly, the results presented here have implications for other ITSs (especially those in ill-defined domains). In particular, our results suggest that researchers and developers should place a greater emphasis on the use of NLP as a means to develop stealth assessments of students' abilities.

5. CONCLUSION

In conclusion, the current study utilized two NLP tools to investigate the potential of NLP techniques to inform stealth assessments of students' reading comprehension skills. Ultimately, such a measure is expected to enhance our student models within iSTART. Overall, the current study suggests that natural language processing techniques can be used to help educators and researchers development stealth assessments and student models within ITSs [39]. These models can then be used to increase the personalization and efficiency of the training these students receive.

6. ACKNOWLEDGEMENTS

This research was supported in part by: IES R305G020018-02, IES R305G040046, IES R305A080589, and NSF REC0241144, NSF IIS-0735682. Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES or NSF. We also thank Scott Crossley, Tanner Jackson, Jennifer Weston, Matt Jacovina, Russell Brandon, Rod Roscoe, and Jianmin Dai for their help with the data collection and analysis and developing the ideas found in this paper.

7. REFERENCES

[1] Murray, T. 1999. Authoring intelligent tutoring systems: An analysis of the state of the art. International Journal of Artificial Intelligence in Education, 10, (1999) 98-129.

[2] Crossley, S. A., Allen, L. K., & McNamara, D. S. 2014. Analyzing discourse processing using a simple natural language processing tool (SiNLP). Discourse Processes, 51, (2014) 511-534.

[3] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. and Louwerse, M. 2004. AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, 36, (2004), 180-193.

[4] McNamara, D. S., Boonthum, C., Levinstein, I. B., and Millis, K. 2007. Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D .S. McNamara, S. Dennis, and W. Kintsch (Eds.), Handbook of Latent Semantic Analysis. Erlbaum, Mahwah, NJ, 227-241.

[5] VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., and Rose, C. P. 2007. When are tutorial dialogues more effective than training? Cognitive Science, 31, (2007), 3-62.

[6] Crossley, S. A., Varner, L. K., Roscoe, R. D., and McNamara, D. S. 2013. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In K. Yacef et al (Eds.), Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED). Springer, Heidelberg, Berlin, 269-278.

[7] McNamara, D. S., Crossley, S. A., and Roscoe, R. D. 2013. Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, 45, (2013) 499-515.

[8] Rus, V., McCarthy, P., Graesser, A. C., and McNamara, D. S. 2009. Identification of sentence-to-sentence relations using a textual entailer. Research on Language and Computation, 7, (2009), 209-229.

[9] Varner, L. K., Jackson, G. T., Snow, E. L., and McNamara, D. S. 2013. Are you committed? Investigating interactions among reading commitment, natural language input, and students' learning outcomes. In S. K. D'Mello, R. A., Calvo, & A. Olney (Eds.), Proceedings of the 6th International

Conference on Educational Data Mining. Springer, Heidelberg, Berlin, 368-369.

[10] Graesser, A. C., Person, N., Harter, D., and the Tutoring Research Group. 2001. Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 12, (2001), 257-279.

[11] Graesser, A. C., Olney, A., Haynes, B.C., and Chipman, P. 2005. AutoTutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In C. Forsythe, M. L. Bernard, T. E. Goldsmith (Eds.), Cognitive systems: Human cognitive models in systems design. Erlbaum, Mahwah, NJ, 177-212.

[12] Graesser, A. C., VanLehn, K., Rose, C., Jordan, P., and Harter, D. 2001. Intelligent tutoring systems with conversational dialogue. AI Magazine, 22, (2001), 39-52.

[13] Rose, C., Roque, A., Bhembe, D., and VanLehn, K. 2002. An efficient incremental architecture for robust interpretation. In Proceedings of Human Languages Technologies Conference. San Diego, CA,

[14] Jordan, P., Makatchev, M., and VanLehn, K. 2004. Combining competing language understanding approaches in an intelligent tutoring system. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), 7th International Conference on Intelligent Tutoring Systems. Springer, Heidelberg, Berlin, 346-357.

[15] McNamara, D. S., Levinstein, I. B., and Boonthum, C. 2004. iSTART: Interactive strategy trainer for active reading and thinking. Behavioral Research Methods, Instruments, & Computers, 36, (2004), 222-233.

[16] Graesser, A. C., McNamara, D. S., and Rus, V. 2007. Computational modeling of discourse and conversation. In M. Spivey, M. Joanisse, & K. McRae (Eds.), Cambridge Handbook of Psycholinguistics. Cambridge University Press, Cambridge, UK.

[17] Graesser, A. C., Jackson, G. T., Mathews, E., Mitchell, H., Olney, A., Ventura, M…& the Tutoring Research Group. 2003. Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Boston, 1-5.

[18] Varner, L. K., Jackson, G. T., Snow, E. L., McNamara, D. S.: Does size matter? Investigating user input at a larger bandwidth. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), Proceedings of the 26th Annual Florida Artificial Intelligence Research Society Conference, AAAI, St. Petersburg, FL, 546-549.

[19] Shute, V. J. 2011. Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), Computer Games and Instruction. Information Age Publishers, Charlotte, NC, 503-524.

[20] Shute, V. J., and Kim, Y. J. 2013. Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology (4th Edition)*. Lawrence Erlbaum Associates, Taylor & Francis Group, New York, NY, 311-323.

[21] Shute, V. J., Ventura, M., Bauer, M. I., and Zapata-Rivera, D. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody & P. Vorderer (Eds.), Serious games: Mechanisms and effects. Routledge, Mahwah, NJ, 295-321.

[22] Snow, E. L., Jacovina, M. E., Allen, L. K., Dai. J., and McNamara, D. S. 2014. *Entropy: A stealth assessment of agency in learning environments.* In J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining*, (London, UK, July 4 -7, 2014), Springer Berlin Heidelberg, 241-244.

[23] McNamara, D. S. 2011. Measuring deep, reflective comprehension and learning strategies: Challenges and successes. Metacognition and Learning, 3, (2011), 1-11.

[24] Varner, L. K., Roscoe, R. D., and McNamara, D. S. 2013. Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. Journal of Writing Research, 5, (2013), 35-59.

[25] Brusilovsky, P. 1994. The construction and application of student models in intelligent tutoring systems. Journal of Computer and Systems Science International, 23, (1994), 70-89.

[26] Vanlehn, K. 2006. The behavior of tutoring systems. International journal of artificial intelligence in education, 16 (2006), 227-265.

[27] Chi, M., Bassok, M., Lewis, M., Reimann, P., and Glaser, R. 1989. Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, (1989), 145-182.

[28] McNamara, D. S. 2004. SERT: Self-explanation reading training. Discourse Processes, 38, (2004), 1-30.

[29] Magliano, J., Todar, S., Millis, K., Wiemer-Hastings, K., Kim, H., and McNamara, D. 2005. Changes in reading strategies as a function of reading training: A comparison of live and computerized training. Journal of Educational Computing Research, 32, (2005), 185-208.

[30] Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. (Eds.). 2007. Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.

[31] Jackson, G. T., Guess, R. H., and McNamara, D. S. 2010. Assessing cognitively complex strategy use in an untrained domain. Topics in Cognitive Science, 2, (2010), 127-137.

[32] Jackson, G. T., Boonthum, C., and McNamara, D. S. 2010. The efficacy of iSTART extended practice: Low ability students catch up. In J. Kay & V. Aleven (Eds.), Proceedings of the 10th International Conference on Intelligent Tutoring Systems. Springer, Heidelberg, Berlin, 349-351.

[33] McNamara, D. S., O'Reilly, T., Best, R., and Ozuru, Y. 2006. Improving adolescent students' reading comprehension with iSTART. Journal of Educational Computing Research, 34, (2006), 147-171.

[34] Bell, C., and McNamara, D. S. 2007. Integrating iSTART into a high school curriculum. In Proceedings of the 29th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Austin, TX, 809-814.

[35] Jackson, G. T., and McNamara, D. S. 2013. Motivation and performance in a game-based intelligent tutoring system. Journal of Educational Psychology, 105, (2013), 1036-1049.

[36] Jackson, G. T., Dempsey, K., and McNamara, D. S. 2010. The evolution of an automated reading strategy tutor: From classroom to a game-enhanced automated system. In M. S. Khine & I. M. Saleh (Eds.), Cognition, Computers and Collaboration in Education, Springer, NY, 283-306.

[37] MacGinitie, W., and MacGinitie, R. 1989.Gates MacGinitie reading tests. Riverside. Chicago, IL.

[38] McNamara, D. S., Graesser, A. C., McCarthy, P., and Cai, Z. 2014. Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press.

[39] Crossley, S. A., Allen, L. K., & McNamara, D. S. 2014. Analyzing discourse processing using a simple natural language processing tool (SiNLP). *Discourse Processes, 51,* 511-534.

.