What'd You Say Again? Recurrence Quantification Analysis as a Method for Analyzing the Dynamics of Discourse in a Reading Strategy Tutor

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

Cecile Perret
Arizona State University
PO Box 872111
Tempe, AZ, 85287
01+404-414-5200
cperret@asu.edu

Aaron Likens
Arizona State University
PO Box 872111
Tempe, AZ, 85287
01+404-414-5200
alikens@asu.edu

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

ABSTRACT

In this study, we investigated the degree to which the cognitive processes in which students engage during reading comprehension could be examined through dynamical analyses of their natural language responses to texts. High school students (n = 142) generated typed self-explanations while reading a science text. They then completed a comprehension test that measured their comprehension at both surface and deep levels. The recurrent patterns of the words in students' self-explanations were first visualized in *recurrence plots*. These visualizations allowed us to qualitatively analyze the different self-explanation processes of skilled and less skilled readers. These recurrence plots then allowed us to calculate recurrence indices, which represented the properties of these temporal word patterns. Results of correlation and regression analyses revealed that these recurrence indices were significantly related to the students' comprehension scores at both surface- and deep levels. Additionally, when combined with summative metrics of word use, these indices were able to account for 32% of the variance in students' overall text comprehension scores. Overall, our results suggest that recurrence quantification analysis can be utilized to guide both qualitative and quantitative assessments of students' comprehension.

LAK '17, March 13-17, 2017, Vancouver, BC, Canada
© 2017 ACM. ISBN 978-1-4503-4870-6/17/03⋯$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027445

Categories and Subject Descriptors

• **Computing methodologies~Natural language processing** • *Applied computing~Computer-assisted instruction* • Applied computing~Psychology

General Terms

Algorithms, Measurement, Performance, Languages, Theory

Keywords

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

1 INTRODUCTION

Literacy is a critically important skill for success in modern society, as individuals are increasingly reliant on text-based communication in their daily lives, classrooms, and workplaces [17; 40]. The ability to learn from and communicate through text relies on an intricate set of processes that include understanding the basic content in the text and generating connections between this new information and prior knowledge of the concepts [32]. Unfortunately, the complexity of these tasks often presents difficulties in students' acquisition of strong literacy skills, as evidenced by consistent reports of low performance on standardized assessments of reading comprehension and writing [37]. Further, teachers often lack the time and resources to provide students the individualized instruction and feedback they need to improve these skills.

In response to this need, researchers have developed educational technologies with the aim of enhancing the quality of the reading and writing training that students receive, as well as their opportunities for deliberate practice (see [10] for an overview). For instance, automated writing evaluation systems deliver automated feedback on students' essay writing [43; 51]. Similarly, Intelligent Tutoring Systems (ITSs) provide students with instruction and automated feedback that can be adapted to their knowledge and skills. For instance, the DSCoVAR (Dynamic

Support of Contextual Vocabulary Acquisition for Reading) system targets students' vocabulary knowledge by providing opportunities to practice reading difficult words across multiple contexts. Additionally, the system provides individualized feedback on students' performance [16].

These literacy-focused adaptive technologies build on a strong foundation of research on the use of artificial intelligence in education. Traditionally, this field has focused on the development and use of ITSs that target instruction in well-defined domains, such as mathematics and physics [36; 38; 47]. The strength of these systems is largely grounded in their ability to adapt the instruction, practice problems, and feedback that students receive based on on-line assessments of their performance, affective states, and knowledge. These systems have been shown to be highly effective, with a recent overview reporting no significant differences in effect size between ITSs and expert one-on-one human tutoring [48].

Despite their obvious similarities, however, educational technologies that target literacy skills (as well as skills in other ill-defined domains) differ from more traditional ITSs in a number of important ways. Perhaps the most salient of these differences is the nature of students' responses to the tutoring system. For example, ITSs that target math instruction can present students with high numbers of multiple-choice questions in a single training session, each of which has a set of right and wrong answers for students to select. Based on the measured performance on these items, the system can adapt additional practice problems and feedback to students' individual needs [47].

Conversely, ITSs for literacy instruction often prompt students to respond to the tutor using *natural language.* For instance, iSTART – an ITS for reading strategy training – prompts students to type self-explanations (i.e., explanations of the meaning of the text material to oneself) of texts as they read [33]. Similarly, We-Write – a tutoring system that provides training on self-regulation strategies for writing – prompts students to generate and revise essays in the system [52]. In the current study, we describe recent work that aims to develop more robust assessments of students' natural language responses in ITSs such as these. In particular, we extend work on discourse analyses by examining the temporal properties of the language that students produce during learning tasks and relate these properties to students' performance at multiple levels.

1.1 Adaptivity in Educational Technology

Educational technologies rely on assessments of student performance to drive adaptive instruction and feedback. In an effort to not distract from the learning process, system developers have increasingly relied on measures that can be collected within the learning task itself [45-46]. These "stealth assessments" can be informed by a wealth of data commonly collected by intelligent tutoring systems, such as the choices students make during learning tasks, the trajectories of their mouse movements, and the keystrokes they press while typing. For example, log data (e.g., students' clicks in the system) has been used to develop detectors of students' engagement [24] and affect [5; 13] during learning tasks.

Once these stealth assessments have been developed, they can be used to develop models of student users. These models then allow the system to individualize the instruction and feedback that students receive based on their strengths and weaknesses [7]. Importantly, these models can be continuously updated in the system as additional data is collected. This ensures that the system is appropriately accounting for changes in students' knowledge and skills over the course of their training in the system.

1.1.1 Language Assessment in Educational Technologies

Albeit still rare, ITS developers increasingly incorporate natural language and natural language processing (NLP) techniques into tutoring systems in an effort to increase adaptivity and learning [18-21; 34; 42]. For instance, Why2 Atlas – a tutoring system for physics – engages students in natural language dialogue related to their qualitative explanations of physics problems [42]. Prior research suggests that these interactions with NLP-based tutoring systems lead to significant learning gains compared to non-interactive learning tasks [18; 49].

More recently, researchers have begun to use the data collected from these natural language responses to develop more nuanced stealth assessments of students' characteristics and performance [3; 12; 34]. For instance, D'Mello and colleagues (2009) [12] found that they could predict the proportional occurrence of students' affective states through analyses of the cohesion in their dialogues with an automated tutor. Similarly, McNamara and colleagues (2007) [34] found that natural language processing indices could be used to accurately score the quality of students' self-explanations during text reading.

Despite this significant progress, NLP-based tutoring systems have plenty of room for improvement. One issue relates to the ability of these systems to measure the *on-line* cognitive and affective processes of student users. Analyses of students' language typically rely on aggregate measures of language use (e.g., the most common words used by students, total number of words produced) and, as such, provide little information about the processes in which students are engaged. In order to provide more nuanced assessments that can target students' needs, ITSs should analyze the properties of students' language as it unfolds over time.

1.1.2 Dynamical Analyses of Natural Language

In the current study, we rely on computational techniques from dynamical systems theory to analyze the temporal organization of students' natural language responses to text. Dynamic methodologies provide a novel means with which researchers can characterize *patterns* that emerge from students' behaviors (e.g., language, system choices) during learning tasks. Traditional statistics often aggregate variables across time, potentially discarding important information about learning and performance. In contrast, dynamic methodologies consider time to be a critical component of the analysis and explicitly seek to characterize temporal patterns. Thus, rather than treating behavior as a static process, these dynamic analyses more accurately account for the complex, changing nature of behavior. Although the current study is one of the first to use dynamic analyses to assess students' natural language responses to an intelligent tutoring system, these techniques have previously been used across a wide variety of domains as a means to understand the complex patterns that manifest in individuals' behaviors over time [e.g., 4; 11; 41; 44].

To illustrate the potentially important value of these dynamic text analyses, consider that you have been asked to read a text on a complex topic and explain the text to yourself as you read. How might the topics you reference change over the course of this task, compared to when you are reading a text more passively? It may be the case that when you read the text passively, you simply explain the meaning of the individual sentences to yourself, without referencing the previous material in the text or your outside knowledge. When you are reading the text more deeply, however, you might read sentences, but consistently refer to previous material in order to generate connections and develop a deeper understanding of the concepts in the text.

The differences described in this example may play an important role in modeling the processes that students are engaging in during text comprehension, which can ultimately help to develop more nuanced assessments of their performance. For instance, it is possible that a students' comprehension of text-based information (i.e., information that does not require the reader to make connections across sentences or paragraphs in the text) can be detected with simple, traditional analyses of the frequent words occurring in their natural language responses to the text. Their deep comprehension of the text (i.e., their performance on items that require the reader to generate inferences), however, may be missed if the temporal nature of these responses is not taken into account. In this scenario, dynamic analyses that account for the temporal distributions of the words in students' text responses may prove more informative than static measures.

1.1.3 Recurrence Quantification Analysis

Here, we utilize a dynamic methodology – recurrence quantification analysis -- to visualize and quantify the extent to which recurrent patterns in students' natural language text responses relate to their reading comprehension processes. Recurrence quantification analysis (RQA) is a nonlinear data analysis technique that provides information about patterns of repeated behavior (i.e., the number and duration of recurrences) in a continuous or categorical time series [30]. Like many techniques used in the dynamical systems theory framework, this methodology has been used in a variety of domains, both within and outside the realm of human behavior [11; 44]. For example, researchers have utilized recurrence quantification analyses to examine patterns of heart-rate variability [30], postural fluctuations [41] and eye movements [4].

Beyond these physiological measures, RQA has the potential to provide important information about recurrence in the content of students' language. Dale and Spivey (2005) [11], for example, have revealed that RQA can be applied to categorical data sets, such as the words in a particular conversation. This flexibility of the RQA technique (i.e., the fact that it can be applied to both continuous and categorical data sets) may be particularly salient for the study of natural language. In particular, recurrence can be measured at multiple levels of the text (e.g., word, semantic), rather than relying only on one level of analysis.

The starting point of RQA is the development of a recurrence plot, which is a visualization of a matrix wherein the individual elements represent points in a time series that are visited more than once (i.e., they recur). In other words, this plot represents the times in which a dynamical system visits the same area in a phase space [29]. Within this plot, each point represents a particular state that is revisited by the system. If multiple points occur continuously, they form diagonal lines, which represent times when the system is revisiting an entire sequence of states.

As a simple illustration, consider the following sentence: "The ice cream man brought ice cream on Friday." To generate a recurrence plot for this sentence, the words in the sentence are first placed on both the X and Y axes of a 2-dimensional plot (see Figure 1). Each time a word appears both the X and Y axes, a dot is placed in that location on the plot. Because this sentence is being plotted against itself, the recurrence plot is symmetrical with a diagonal line through the center – the line of identity (LOI). The points of interest in these recurrence plots are the points that do not occur on the main diagonal. Individual points off the main diagonal represent the times that a word is repeated later in the sentence. When multiple points occur simultaneously, these points form *diagonal lines* (e.g., "ice cream" in Figure 1)*,* which represent *sequences* of words that are repeated in time.

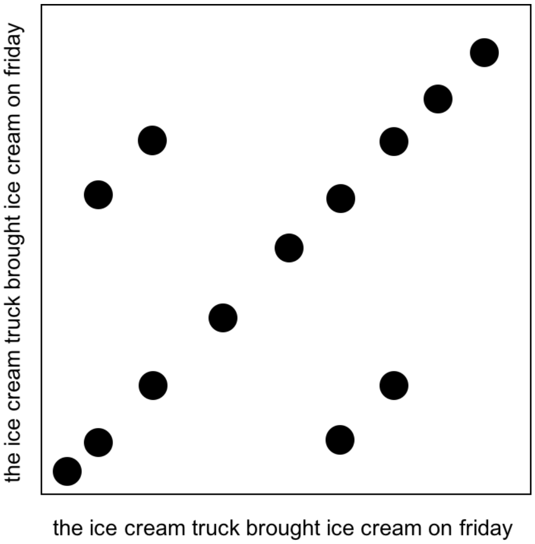

Figure 1. Example recurrence plot

Visualizing recurrent patterns is informative, but researchers also need to quantify the structure contained in recurrence plots. Recurrence quantification analysis offers multiple metrics that help to quantify recurrent patterns to allow for statistical comparisons of recurrence plots [53]. Below, we briefly describe the most commonly used metrics in recurrence quantification analyses. For more detailed information, see [9].

Recurrence Rate. The recurrence rate is a measure of the density of points represented in a recurrence plot. A recurrence plot is calculated by dividing the total number of points in a plot by the square of the length of the overall time series. This metric represents the overall amount of recurrence that is present in the recurrence plot, regardless of the distributions of the points.

Determinism. Determinism is a measure of the number of recurrent points that tend to fall on diagonal lines (ignoring the LOI) in the recurrence plot. Thus, this metric provides information about the *distribution* of the recurrent points. Diagonal lines in recurrence plots reflect time periods when the system is revisiting a particular sequence of states. Thus, systems with low determinism can exhibit short moments of repetitive states; however, they are considered less ordered than highly deterministic systems.

Average Line Length. This metric calculates the average length of the diagonal lines in the recurrence plot. Thus, when the system repeats a sequence of states, this metric provides information about the typical length of those sequences.

Maximum Line Length. This metric calculates the length of the longest diagonal line in the recurrence plot. Therefore, this metric reveals whether a system revisits a long sequence of states at some point in time.

Entropy. Entropy is calculated as the Shannon entropy of the distribution of the line lengths in the recurrence plot. This metric quantifies the degree to which the trajectory of the system exhibits order. Thus, entropy will be higher if the system revisits a wider variety of state sequences over time. Dynamic systems that continually revisit the same, or similar, sequences of states, will have lower entropy.

1.2 iSTART

This study aims to refine the adaptive capabilities of the Interactive Strategy Training for Active Reading and Thinking (iSTART) system, an intelligent tutoring system that teaches high school and college students self-explanation strategies to improve their comprehension of complex texts [28; 33]. Self-explanation has repeatedly been shown to be beneficial for improving higher order skills such as deep comprehension of text, inference generation, and problem solving [8]. In this study, we intend to maximize iSTART's ability to produce a user model in order to improve the system's adaptability to individual student's needs.

iSTART is based on the Self-Explaining and Reading Training (SERT) intervention, which was created to teach students effective strategies for self-explaining a text followed by practice on how to use them as they read [31]. Previous research has demonstrated the effectiveness of SERT, as well as iSTART, in improving students reading comprehension skills of complex texts [28]. iSTART focuses on self-explanation training through two principal modules within the system: training and practice.

In the training module, students are taught the self-explanation strategies (comprehension-monitoring, paraphrasing, prediction, elaboration, and bridging) through the use of animated videos presented by a pedagogical agent. Each strategy is taught through the use of definitions, mnemonic devices, and examples. Students then answer a set of checkpoint questions to determine their comprehension of the lesson. Once all students have received a 75% score on each of the lesson checkpoints, they view a summary lesson and are then prompted to practice using the strategies in an initial practice activity. In this phase, students use the self-explanation strategies and receive feedback for two texts. Once completed, students are ushered into the practice module of the system.

The practice module is composed of a variety of practice activities, all falling in one of two categories: identification and generative. Identification mini-games are all games in which students are prompted to read a text with an associated self-explanation and then must select the specific strategy that was used for that self-explanation. Generative practice, however, is composed of both game-based practice as well as non-game practice activities. These activities prompt students to generate their own self-explanations to a designated target sentence as they read a text. Students are then given feedback on their response based on a complex algorithm that uses linguistic indices to determine the quality of the student's response. Generative practice activities are designed to allow teachers the opportunity to insert their own texts for their students, thus the evaluation algorithm that iSTART uses to score the self-explanations must be flexible and accurate to optimize the system's capabilities to improve student learning outcomes.

1.2.1 iSTART Evaluation Algorithm

iSTART is designed to assess and score students' self-explanations immediately after each individual submission to the system. Since analyses are always conducted on a local basis, the system is dependent on a limited set of available information. This includes the student's response, the specific target sentence prompting the self-explanation, and the previous sentences of the text. The algorithm uses both word-based indices and Latent Semantic Analyses (LSA) to assess the quality of the self-explanation and determine the appropriate feedback. Lower-level assessments are informed by word-based indices that include response length as well as quantity of content-word overlap. Typically, these provide an initial report on whether the response is too short, too similar or identical to the topic sentence, or entirely irrelevant. After these initial analyses, more information is taken into consideration using LSA, which is capable of producing a more holistic assessment by determining how well the self-explanation is related to the text as well as outside content (considered as a student's prior-knowledge).

The algorithm produces a score using word-based indices and LSA-indices on a scale from 0 to 3. A score of 0 is given when the self-explanation is either too short, irrelevant, or too similar to the target sentence. A score of 1 demonstrates that the student wrote a self-explanation that solely relates to the target sentence. A score of 2 is generated if the student wrote a self-explanation that relates to both the target sentence and previous portions of the text. Students receive a score of 3 when their self-explanations derive information from the target sentence, previous sentences of the text, as well as external information not directly related, though relevant, to the text. This implies that the student not only produced inferences throughout the text, but elaborated on the available information using background knowledge. Research using the iSTART algorithm has shown that it scores as accurately as humans and that it can offer a summary of the cognitive processes used in reading comprehension [25].

1.2.2 Aggregated Self-Explanation Analyses

Previous research on the iSTART system has relied on NLP techniques to analyze student's self-explanation responses [1-3; 50]. Initial work focused on local sentence-level analyses of students' individual self-explanations to texts. However, recent research has begun to observe how a set of self-explanations that span an entire text can be aggregated and evaluated to provide analyses at a more global level. Such analyses reveal a far more comprehensive interpretation of the comprehension processes involved in reading a text.

Research on these "aggregated self-explanations" was motivated by the possibility of increasing the bandwidth of available information to analyze. Researchers used analyses of aggregated self-explanations to determine whether evaluating responses at a larger window size (i.e., aggregated self-explanations for a single text as opposed to individual self-explanations) would improve upon a student model [3; 50]. Results showed that analyzing the aggregated responses accounted for an additional 10% of the variance that was already accounted for by the original iSTART algorithm [50]. These studies also show a positive relationship between the aggregated NLP scores and iSTART algorithm scores and pretest reading scores [3].

Additional research studies have assessed which specific linguistic indices provide more accurate predictions of potential connections within the text. Studies have determined that indices relating to local and global cohesion are most likely to reveal relevant connections being made across self-explanations. Specifically, when students' aggregated self-explanations display a higher incidence of causal cohesion, these students also exhibit better comprehension of the text [1-2].

Recently, Allen, Jacovina, and McNamara (2016) [1] discovered that over the course of multiple sessions of practice within iSTART, the global cohesion of students' aggregated self-explanations increased. This implies that over time students learn to generate more inferences and create deeper connections across the text they read. This finding demonstrates that extended training within iSTART improves student comprehension processes. Ultimately, these linguistic signatures can provide the system with information on students' performance over time, thus helping to determine what type of practice is optimal for individualized training.

1.3 Current Study

The purpose of the current study is to investigate the degree to which the cognitive processes in which students engage during reading comprehension can be examined through dynamical analyses of their natural language responses to the text. We use dynamic *visualizations* and *quantifications* of students' natural language text responses to measure their performance on a comprehension test. In particular, we examine whether the patterns of students' word usage during their text responses reflect differences in their cognitive processes, as reflected by their performance on surface- and deep-level comprehension questions. Additionally, we present visualizations of these patterns and provide qualitative assessments of these visualizations to demonstrate their potential to drive student feedback.

2 METHODS

2.1 Participants

The data for this study was collected as part of a larger, five-session study. In total, 149 high school and college freshmen (6.7%) students participated in this study located in the southwestern United States. On average, the students were 15.69 years of age (range = 13-19). Of these students, 55% were female and 16.8% reported speaking English as a second language Additionally, 43.6% were Caucasian, 32.2 %were Hispanic, 8.7 % were African-American, 7.4 % were Asian, and 8.1 % reported other nationalities. Seven students were dropped from the analyses due to data loss and attrition; thus, we analyzed data for 142 total students.

2.2 Study Procedure

The data included in this study was collected over the course of two sessions, which lasted between one and two hours. In the first session, students' general world knowledge, reading comprehension and writing skills, and attitudes were assessed using the following measures: Demographics questionnaire, Alternate Uses task [22]; selected items from the Remote Associative task (RAT) [35]; Motivated Strategies for Learning Questionnaire (MSLQ) [39]; Gates MacGinitie Reading test (Gates-MacGinitie (4th ed.) reading skill test (form S) level 10/12) [27]; 30 question multiple-choice test on general knowledge in literature, science, and history; 25 minute timed-essay writing task; and a Component Processes test [23].

The data collected in session two was collected one to three days after session one and contained the following measures: Cognitive Reflection Test (CRT) [15]; Self-Explanation and Reading Comprehension Test; On-line Motivation Questions [6]; Learning Orientation and Performance Orientation task (LO/PO) [26]; and a Grit assessment [14].

For the purposes of the current study, we only analyzed the data from the Demographics questionnaire in session one and the Self-Explanation and Reading Comprehension test in session two.

2.3 Self-Explanation and Reading Comprehension Test

A Self-explanation and Reading Comprehension Test was administered to students to analyze the on-line reading processes students employed during reading, as well as their comprehension of the text at the surface (text-based) and deep (bridging) levels. Students read and self-explained one of two science texts during session two related to heart disease or red blood cells. This text was presented one segment (i.e., two to three sentences) at a time, with each segment separated by a target sentence in bold. For each target sentence, students were instructed to write a self-explanation of the information they had just read. In total, each student wrote nine self-explanations for the text.

Immediately following this self-explanation and reading procedure, the students were asked to answer eight comprehension questions. The comprehension test consisted of 4 text-based and 4 bridging open-ended questions. The text was not visible to students while they answered these questions. Text-based questions were based on information found within one sentence in the text, whereas bridging questions required students to refer to information from two or more sentences within the text. Each question was worth one point, but allowed partial credit. Thus, the maximum number of points that a student could receive on this test was eight. The comprehension questions were independently scored by two expert raters for at least 14% of the responses. Raters resolved discrepancies and repeated the process until they received 95% exact agreement, with a kappa of at least 0.8. Once interrater reliability was achieved, one coder completed the remainder of the scoring.

Table 1. Recurrence Quantification Analysis Indices

	Description
Recurrence Rate	Proportion of the recurrence plot that is composed of recurrent points
Determinism	Proportion of recurrent points that form diagonal line structures (defined as 2 or more recurrent points in a row)
Line Number	Total number of lines in the recurrence plot.
Max Line	Length of the longest diagonal line in the plot, excluding the main diagonal
Average Line	Average length of the lines in the recurrence plot
Entropy	Shannon information entropy of diagonal line lengths
Normalized Entropy	Entropy variable normalized by the number of lines in the plot

2.4 Data Processing

For the purpose of generating and quantifying the recurrence plots, students' individual, sentence-level self-explanations were aggregated. Therefore, each student had one "aggregated self-explanation" file that included the nine self-explanations they produced while reading.

To prepare the data for the RQA, the texts were first cleaned. All punctuation in the texts was first removed and the words were all converted to lower case and stemmed. Once the texts were cleaned, the series of words was converted to series of categorical numeric codes, which each represented the unique words in each self-explanation. For instance, the sentence, "The bird ate bird food." would be converted to the series: {1, 2, 3, 2, 4}.

2.5 Recurrence Quantification Analyses

We used the *crqa* library in R [9] to generate the recurrence plots and calculate the recurrence indices for students' self-explanations. The resulting indices are described in Table 1.

2.6 Text Analyses

In addition to the RQA indices, descriptive indices of students' aggregated self-explanations were calculated to provide summary information about the words in students' self-explanations. Specifically, we calculated the *total number of words*, the *number of letters per word*, and the *type-token ratio.* The type-token ratio is a measure of the number of unique words in the text divided by the total number of words. We included these basic text indices in our analysis to determine whether the recurrence quantification analysis metrics accounted for different and unique variance beyond these basic descriptive indices.

2.7 Statistical Analyses

To assess the degree to which the patterns of recurrence in students' self-explanations were associated with their comprehension of the text, we generated recurrence plots and calculated Pearson correlations and regression analyses. The recurrence plots allowed us to *visualize* the recurrent word patterns across students' self-explanations of the text. Additionally, these recurrence plots allowed us to *quantify* the properties of these plots with seven RQA indices (see Table 1).

Normality of the indices was assessed with skew, kurtosis, and visual data inspections. Two indices, *Line Number* and *Average Line* were strongly skewed; therefore, we calculated the log transformation for this index.

Pearson correlations were used to assess relations between word recurrence (as defined by the RQA indices) and comprehension scores. We calculated these correlations for students' overall comprehension scores, as well as their text-based and bridging comprehension scores. Finally, stepwise regression analyses were conducted to follow-up the correlation analyses in order to provide an indication of the variables that accounted for the most variability in the dependent variables. For this analysis, we included the three basic descriptive indices and the RQA indices to determine whether the RQA indices accounted for unique variance in the model once the basic indices were included. Multicollinearity was assessed among the indices ($r > .90$) included in the regression analysis; however, no indices demonstrated multicollinearity. Additionally, the self-explanations of eleven students contained fewer than 100 words, which did not provide enough data points for the Entropy RQA indices to be calculated. Therefore, we conducted pairwise deletion to account for this missing data in our correlation and regression analyses.

3 RESULTS

3.1 Qualitative Analysis of Recurrence Plots

To visualize the temporal distribution of words in students' self-explanations, recurrence plots for each student were calculated using the procedure described in the previous sections. These recurrence plots varied considerably among the students and provided us a means to *qualitatively* analyze differences in the word recurrence in the self-explanations of students who received low and high scores on the comprehension test.

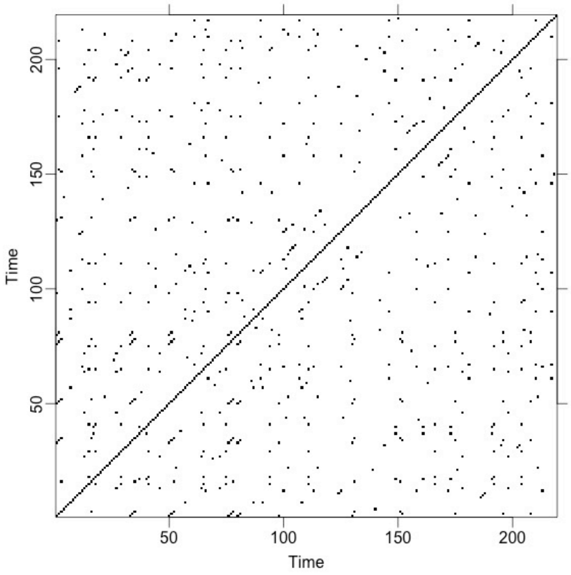

Figure 2. Recurrence Plot for a Student with a Low Text Comprehension Score

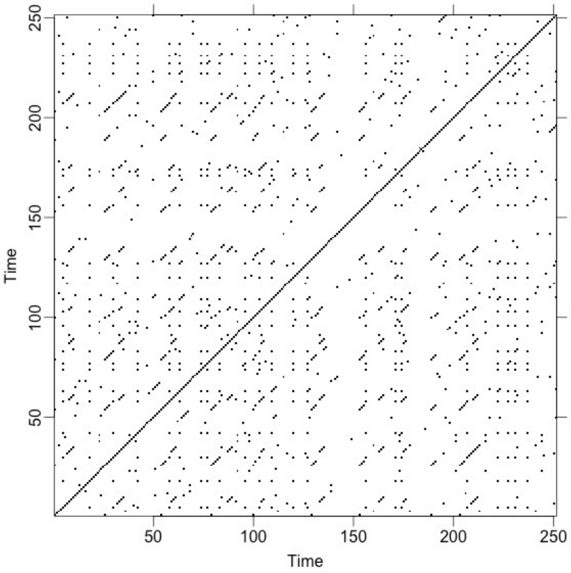

Figure 3. Recurrence Plot for a Student with a High Text Comprehension Score

Figures 2 and 3 illustrate two recurrence plots that were generated using two students' actual self-explanations from the current study. Although the students' self-explanations had a similar total number of words (Figure 2 = 224; Figure 3 = 251), the plots demonstrate that these students exhibited strongly different patterns of word recurrence throughout their self-explanations.

Figure 2 illustrates the recurrence plot of a student who received a score of 1 (out of 8) on the comprehension test (text-based comprehension score = 1; bridging comprehension score = 0). As can be seen in the plot, this student rarely produced self-explanations with similar words from their previous explanations. Additionally, in the situations when this student did exhibit word recurrence, the words tended to occur in isolation, rather than in sequences (diagonal lines) of words. In other words, the recurrence plot suggests that this student did not generate explicit connections between the information explained in different sections of the text.

In contrast, the plot depicted in Figure 3 comes from a student who received a perfect score of 8 on the comprehension test (text-based comprehension score = 4; bridging comprehension score = 4). Unlike the previous student, this student exhibited a high degree of recurrence across self-explanations. Additionally, many of the recurrent points fell on diagonal lines, suggesting that this student was repeatedly referring to sequences of words, rather than individual words. Thus, while reading through the text, the student continued to explain the new text information in connection with previously encountered text information.

Overall, these recurrence plots provide a means through which the comprehension processes of skilled and less skilled readers can be differentiated. Despite the fact that these two students generated a similar amount of text during the self-explanation procedure, the temporal distribution of the words they used varied widely. In particular, these plots reveal that the student who continuously repeated words and phrases while self-explaining ultimately developed a deeper comprehension of the text. In comparison, the student who rarely repeated information across self-explanations demonstrated low text comprehension.

3.2 Text Comprehension

The qualitative analyses of the recurrence plots provided preliminary evidence that skilled and less skilled readers exhibited strong differences in their word recurrence during self-explanation. To empirically test these findings, we conducted quantitative analyses of these plots.

Table 3. Correlations between RQA indices and Comprehension Scores

RQA Index	Text-Based	Bridging	Total
Recurrence Rate	.119 (M)	.101	.126 (M)
Determinism	-.058	.071	.001
Log of Line Number	.413**	.481**	.505**
Max Line	.132 (M)	.142*	.155*
Log of Average Line	.002	.177*	.093
Entropy	.011	.229*	.124 (M)
Normalized Entropy	-.204*	.019	-.116

p <.001**; *p* <.05*; Marginal = M

Pearson correlations were first calculated between the RQA indices and students' text comprehension scores (see Table 3). Results from these analyses indicated that students' comprehension scores were significantly related to a number of the RQA indices. In particular, these results reveal that skilled readers did not simply repeat words more often than less skilled readers. Rather, they differed from less skilled readers in their more frequent repetition of longer sequences of words. Importantly, the relations between the RQA indices and comprehension scores differed between text-based and bridging questions. These findings suggest that these recurrence characteristics are able to provide nuanced information about students' comprehension processes that go beyond holistic comprehension scores.

We conducted three stepwise regression analyses with the RQA indices and three basic text indices (i.e., *total number of words*, the *number of letters per word*, and the *type-token ratio*) as predictors and the comprehension scores (i.e., total, text-based, and bridging) as the dependent variables. The purpose of these analyses was to assess the amount of variance accounted for by the RQA indices, as well as to determine whether these indices accounted for variance in the comprehension scores when summative text measures were taken into account.

The three regression analyses yielded significant models. The analysis of students' total comprehension scores [*F* (2, 118) = 27.58, p < .001; R^2 = .32] retained two variables: Log of Line Number [β = .54, *p* < .001] and Number of Letters per Word [β = .25, *p* < .01].

The analysis of students' text-based comprehension scores [*F* (3, 117) = 11.60, p < .001; R^2 = .23] retained three variables: Log of Line Number [β = .48, *p* < .001], Number of Letters per Word [β = .19, *p* < .05], and Determinism [β = -.18, *p* < .05].

Finally, the analysis of students' bridging comprehension scores [*F* (4, 116) = 11.18, p < .001; R^2 = .38] retained four variables: Log of Line Number [β = .70, *p* < .001], Number of Letters per Word [β = .25, *p* < .01], Normalized Entropy [β = .32, *p* < .01] and Determinism [β = -.26, *p* < .05].

The results of these analyses suggest that students' text comprehension was most strongly predicted by the number of diagonal lines in their recurrence plots, as well as the size of their words. This provides confirmation of the qualitative analyses by indicating that the skilled readers more frequently repeated sequences of words, rather than individual words. In addition, the words that skilled readers use tend to be longer, or less frequent words, which provides a proxy for students' vocabulary. Additionally, the analyses revealed that the recurrence metrics were more strongly related to students' performance on bridging questions than text-based questions. Thus, comprehension questions that required students to generate connections across multiple sentences in the text were more strongly related to the word recurrence in students' self-explanations.

4 DISCUSSION

Educational technologies across a variety of domains increasingly incorporate natural language components for the purpose of increasing interactivity and providing students with adaptive instruction and feedback [10; 20; 42]. While these systems generally provide accurate holistic feedback [34; 43; 51], they often lack the more nuanced information that is needed to drive formative feedback related to beneficial learning processes. The objective of many natural language assessments is to deliver accurate scores that match an expert's ratings of quality. However, the indices used in these analyses often exist in a "black box" and can be difficult to translate into actionable feedback for students. Additionally, these assessments do not often take the temporal aspects of language into account, which may play a critical role in the assessment of students' performance at more fine-grained sizes.

In this study, we addressed these research gaps through computational analyses of students' natural language responses to a text. We leveraged dynamic modeling techniques to capture the temporal properties of students' language use and to relate those properties to students' performance on a comprehension test. Importantly, this analysis did not solely rely on statistical assessments of student performance. We were able to generate metrics that could provide both qualitative and quantitative information about students' comprehension performance. We anticipate that these metrics will be able to drive summative feedback in educational technologies, but also provide students with meaningful visualizations of their work. These visualizations may ultimately help students to ground the system feedback in specific examples from their own work, which can lead to improvements in their understanding and uptake of the feedback.

The results of the current study support our hypotheses that the temporal, recurrent properties of students' text responses can provide important information about their comprehension. The qualitative analyses of students' recurrence plots indicated that successful comprehension processes could be observed through visualizations of students' word use over time. Specifically, the skilled reader depicted in Figure 3 consistently repeated sequences of words across self-explanations, whereas the less skilled reader (Figure 2) referred to previously mentioned concepts much less frequently. This is an important finding because it indicates that the temporal variability in students' natural language responses can provide important information about their comprehension processes. Further, these analyses revealed that visualizations of these language sequences can be used to deliver meaningful information about these different comprehension processes.

The RQA indices generated from these recurrence plots were additionally able to provide important information about students' comprehension performance. In particular, the results of the correlation and regression analyses indicated that 32% of the variance in students' comprehension scores were accounted for using a combination of summative metrics of word use (i.e., *total number of words*, the *number of letters per word*, and the *type-token ratio*), as well as indices related to recurrent patterns of this word use. These analyses speak to the importance of accounting for temporal patterns in analyses of students' language. Natural language processing techniques tend to rely on summative metrics of text features; however, the results of the current study suggest that expanding these analyses to include temporality can provide critical information about students' learning processes.

The correlation analyses additionally revealed similarities and differences between the relationships between these recurrence metrics and the text-based and bridging comprehension scores. Performance on both the text-based and bridging questions was related to a greater number of recurrent word sequences (Log of Line Number) and a longer maximum recurrent sequence (Max Line). This is an interesting finding and suggests that comprehension at multiple levels can be enhanced through the generation of connections among text information. In particular, both text-based and bridging scores demonstrated medium relationships to the number of lines in students' recurrence plots. Thus, feedback driven by these metrics could potentially be developed to prompt students to generate greater connections among ideas in order to improve their understanding of the text content.

In addition to this similarity in recurrent lines, the correlations were indicative of some interesting differences between the text-based and bridging scores. For instance, while bridging scores were positively associated with the raw entropy index for the line lengths in students' plots, text-based comprehension scores were negatively related to the normalized entropy metric. This suggests that the processes underlying students' surface- and deep-level comprehension performance may differentially manifest in the temporal properties of their response to texts. This has important implications for future system adaptability. If these findings were to be replicated with more descriptive information in follow-up studies, it suggests that text-based and bridging comprehension performance could be assessed and, therefore, addressed in different ways through system feedback.

As a final note, in the current study, we only focused on the individual words in students' self-explanations, and did not account for the numerous properties that can be calculated in linguistic analyses. This methodological choice was made to provide a demonstration of the power of the recurrence quantification technique when only words are considered. In reality, however, this technique is highly flexible and can be used to analyze any number of features of language. For instance, categorical recurrence quantification analyses (such as this one) can be used to analyze recurrent patterns in the parts-of-speech or topics of students' language. Additionally, recurrence quantification analyses can be applied to model continuous data, such as word frequency or similarity to the topic. Future studies should be conducted to build on the results of the current study to account for the multi-dimensional properties of the language that students generate.

Overall, our results suggest that recurrence quantification analysis can be utilized to guide both qualitative and quantitative assessments of students' comprehension. Our eventual goal is to use these indices to develop more nuanced stealth assessments and formative feedback in the iSTART system. More broadly, the current study suggests that dynamic visualizations and analyses can be used as a step towards more adaptive educational technologies for literacy, as well as for any system that collects students' natural language responses. Although this is only a first step, and a number of studies remain to be conducted, this study provides a strong initial foundation because it demonstrates the feasibility of such measures for modeling student performance.

5 ACKNOWLEDGMENTS

This research was supported in part by: IES R305G020018-02, IES R305G040046, IES R305A080589, and NSF REC0241133, and NSF IIS-0735682. Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES or NSF. We also thank Matt Jacovina, Scott Crossley, Rod Roscoe and Jianmin Dai for their help with the data collection and developing the ideas found in this paper.

6 REFERENCES

[1] Allen, L. K., Jacovina, M. E., and McNamara, D. S. 2016. Cohesive features of deep text comprehension processes. In J. Trueswell, A. Papafragou, D. Grodner, and D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society in Philadelphia, PA,* 2681-2686. Austin, TX: Cognitive Science Society.

[2] Allen, L. K., McNamara, D. S., and McCrudden, M. T. 2015. Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, and P. Maglio, (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (Cog Sci 2015)*, 78-83. Pasadena, CA: Cognitive Science Society.

[3] Allen, L. K., Snow, E. L., and McNamara, D. S. 2015. Are you reading my mind? Modeling students' reading comprehension skills with Natural Language Processing techniques. In J. Baron, G. Lynch, N. Maziarz, P. Blikstein, A. Merceron, and G. Siemens (Eds.), *Proceedings of the 5th International Learning Analytics & Knowledge Conference (LAK'15)*, 246-254. Poughkeepsie, NY: ACM.

[4] Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F. and Kingstone, A. 2013. Recurrence quantification analysis of eye movements. *Behavior research methods*, *45*(3), 842-856.

[5] Baker, R., and Ocumpaugh, J. 2015. Interaction-based affect detection in educational software. In R. Calvo, S. D'Mello, J. Gratch & A. Kappas (Eds.), *The Oxford handbook of affective computing*, 233-245. New York: Oxford University Press.

[6] Boekaerts, M., 2002. The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. *Advances in motivation and achievement*, *12*, 77-120.

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

[8] 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*, 145-182.

[9] Coco, M. I. and Dale, R. 2013. Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *arXiv preprint arXiv:1310.0201*.

[10] Crossley, S. A. and McNamara, D. S. (Eds.). 2016. Adaptive educational technologies for literacy instruction. New York: Taylor & Francis, Routledge.

[11] Dale, R. and Spivey, M. J., 2005. Categorical recurrence analysis of child language. In *Proceedings of the 27th annual meeting of the cognitive science society*, 530-535. Mahwah, NJ: Lawrence Erlbaum.

[12] D'Mello, S., Dowell, N., and Graesser, A. 2009. Cohesion relationships in tutorial dialogue as predictors of affective states. In Dimitrova V., Mizoguchi R., du Boulay B., Graesser A. (eds.) *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 9–16. IOS Press, Amsterdam.

[13] D'Mello, S. and Graesser, A. 2015. Feeling, thinking, and computing with affect-aware learning technologies. In R. Calvo, S. D'Mello, J. Gratch & A. Kappas (Eds.), *The Oxford handbook of affective computing*, 419-434. New York: Oxford University Press.

[14] Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. 2007. Grit: Perseverance and passion for long-term goals. *Journal of personality and social psychology*, *92*(6), 1087-1101.

[15] Frederick, S. 2005. Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25-42.

[16] Frishkoff, G. A., Collins-Thompson, K., Hodges, L., and Crossley, S., 2016. Accuracy feedback improves word learning from context: evidence from a meaning-generation task. *Reading and Writing*, *29*(4), 609-632.

[17] Geiser, S. and Studley, R. 2001. UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. Oakland, CA: University of California.

[18] Graesser, A. C., Chipman, P., King, B., McDaniel, B., and D'Mello, S. 2007. Emotions and learning with auto tutor. Frontiers in Artificial Intelligence and Applications, 158, 569-571.

[19] Graesser, A. C., Chipman, P., Haynes, B. C. and Olney, A. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612-618.

[20] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., and Louwerse, M. M. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 180-192.

[21] Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W. and Harter, D. 2001. Intelligent tutoring systems with conversational dialogue. *AI magazine, 22*(4), 39-51.

[22] Guilford, J. P., Christensen, P. R., Merrifield, P. R., and Wilson, R. C. 1978. Alternate uses: Manual of instructions and interpretation. *Orange, CA: Sheridan Psychological Services*.

[23] Hannon, B. and Daneman, M. 2001. A new tool for measuring and understanding the individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93,* 103-128.

[24] Haswell, R. H. 2006. Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In: P. F. Ericsson and R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences,* 57–78. Logan, UT: Utah State University Press.

[25] 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,* 127-137.

[26] Jha, S. and Bhattacharyya, S. S. 2013. Learning orientation and performance orientation: scale development and its relationship with performance. *Global Business Review*, *14*(1), 43-54.

[27] MacGinitie, W. H. and MacGinitie, R. K. 1989. *Gates MacGinitie reading tests*. Chicago, IL: Riverside.

[28] 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*, 185-208.

[29] Marwan, N., Romano, M. C., Thiel, M., and Kurths, J., 2007. Recurrence plots for the analysis of complex systems. *Physics reports*, *438*(5), 237-329.

[30] Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., and Kurths, J. 2002. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *Physical review E*, *66*(2), 1-8.

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

[32] McNamara, D. S. and Magliano, J. P. 2009. Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation*. New York, NY: Elsevier Science.

[33] 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*, 222-233.

[34] 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*, 227-241. Mahwah, NJ: Erlbaum.

[35] Mednick, S., 1962. The associative basis of the creative process. *Psychological review*, *69*(3), 220-232.

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

[37] National Assessment of Educational Progress. 2011. *The nation's report card: Writing 2011*. Retrieved Nov. 5, 2012, nces.ed.gov/nationsreportcard/writing.

[38] Nkambou, R., Mizoguchi, R., and Bourdeau, J. (Eds.). 2010. *Advances in intelligent tutoring systems* (Vol. 308). Springer Science & Business Media.

[39] Pintrich, P. R. and De Groot, E. V. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, *82*(1), 33-40.

[40] Powell, P. 2009. Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 66,* 664-682.

[41] Riley, M. A., Balasubramaniam, R., and Turvey, M. T., 1999. Recurrence quantification analysis of postural fluctuations. *Gait & posture*, *9*(1), 65-78.

[42] Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K. and Weinstein, A. 2001. Interactive conceptual tutoring in Atlas-Andes. In *Proceedings of AI in Education 2001 Conference*, 151-153.

[43] Shermis, M. and Burstein, J. (Eds.). 2003. *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Erlbaum.

[44] Shockley, K., Santana, M. V., and Fowler, C. A., 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326-332.

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

[46] 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)*, 311-323. New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

[47] VanLehn, K. 2006. The behavior of tutoring systems. International Journal of Artificial Intelligence in Education, 16, 227-265.

[48] VanLehn, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46,* 197-221.

[49] 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*, 3-62.

[50] Varner, L. K., Jackson, G. T., Snow, E. L., and McNamara, D. S. 2013. Does size matter? Investigating user input at a larger bandwidth. In C. Boonthum-Denecke and G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 546-549. Menlo Park, CA: AAAI Press.

[51] Warschauer, M., and Ware, P. 2006. Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*, 1–24.

[52] Wijekumar, K. K., Harris, K. R., Graham, S., and Meyer, B. J. F. 2016. We-Write. In S. A. Crossley and D. S. McNamara (Eds.) *Adaptive educational technologies for literacy instruction*, 184-203. New York: Taylor & Francis, Routledge.

[53] Zbilut, J. P. and Webber, C. L., 1992. Embeddings and delays as derived from quantification of recurrence plots. *Physics letters A*, *171*, 3-4, 199-203.