Recurrence Quantification Analysis:
A Technique for the Dynamical Analysis of Student Writing

Laura K. Allen, Aaron D. Likens, and Danielle S. McNamara

Arizona State University, PO Box 872111, Tempe, AZ, 85287
LauraKAllen@asu.edu, alikens@asu.edu, Danielle.McNamara@asu.edu

Abstract

The current study examined the degree to which the quality and characteristics of students' essays could be modeled through dynamic natural language processing analyses. Undergraduate students (n = 131) wrote timed, persuasive essays in response to an argumentative writing prompt. Recurrent patterns of the words in the essays were then analyzed using recurrence quantification analysis (RQA). Results of correlation and regression analyses revealed that the RQA indices were significantly related to the quality of students' essays, at both holistic and sub-scale levels (e.g., organization, cohesion). Additionally, these indices were able to account for between 11% and 43% of the variance in students' holistic and sub-scale essay scores. Overall, our results suggest that dynamic techniques can be used to improve natural language processing assessments of student essays.

Introduction

Adaptive educational technologies aim to improve student learning by relieving some of the pressures faced by instructors, as well as by providing students with personalized practice opportunities (Crossley & McNamara, 2016; Nkambou, Mizoguchi, & Bourdeau, 2010). These technologies increasingly rely on natural language processing (NLP) techniques to extract information about student performance and individual differences (Allen, Snow, & McNamara, 2015; Graesser, Chipman, Haynes, & Olney, 2005; McNamara, Boonthum, Levinstein, & Millis, 2007). Compared to non-interactive learning tasks, these NLP-based tutoring systems have been shown to lead to significant gains in student learning (e.g., Graesser et al., 2005).

A principal strength of the NLP techniques employed by these systems is their calculation of linguistic information across a wide variety of dimensions and window sizes (e.g., word sophistication, sentence complexity, document cohesion). For example, analyses conducted at the level of individual words can reveal information about the topics (Blei, Ng, & Jordan, 2003) and concreteness of the language found in a document (Brysbaert, Warriner, & Kuperman, 2014). Similar analyses can be conducted at the sentence, paragraph, and document levels to reveal information about student writing, such as lexical sophistication and cohesion (see McNamara et al., 2014 for a review).

Importantly, once these indices have been calculated, they can be used to model information about students' performance on learning tasks or individual differences in their knowledge and skills. For instance, researchers have used these indices to predict expert ratings of essay quality (e.g., Dikli, 2006; McNamara et al., 2015; Shermis & Burstein, 2003) as well as individual differences in their reading skills (Allen, Snow, & McNamara, 2015) and affective states (Allen et al., 2016; D'Mello, Dowell, & Graesser, 2009). Overall, extensive prior research indicates that these NLP techniques can produce powerful sources of data for the development of educational assessments and adaptive educational technologies.

Despite their success, however, these NLP techniques have substantial room for improvement. One particularly salient weakness of these techniques relates to the fact that the majority of indices are calculated based on aggregate metrics of student language. For example, the *lexical sophistication* of a student's essay would be calculated as an average value of the sophistication of all words produced by the student, but would not take into consideration how these words were distributed throughout the essay. As a consequence, these indices may miss out on important nuances in the structure of student writing.

Dynamic Language Analyses

In the current study, we address this gap by conducting dynamic computational analyses of the words in students' essays. Analyses of the dynamic patterns in students' language provide a method through which researchers can model the ways in which language is structured. NLP analyses traditionally calculate aggregate measures of linguistic features over time, which potentially miss out on important information about language structure. Dynamic techniques, on the other hand, consider time to be of critical importance and intentionally factor temporal patterns into analyses. In this way, dynamic methodologies more appropriately account for the complexity that is inherent in language as it unfolds over time. Although not commonly applied to language, dynamical techniques have been previously used in a variety of scientific domains as a means of characterizing human behavior (e.g., Anderson, Bischof, Laidlaw, Risko, & Kingstone, 2013; Dale & Spivey, 2005; Shockley, Santana, & Fowler, 2003).

To illustrate the purpose of these dynamic language analyses, consider a student, Josie, who has been asked to write a persuasive essay that responds to the question: *Do people achieve more success by cooperation or by competition?* How might the words that Josie uses change over the course of the essay that she produces? If Josie has less knowledge of the topic (and consequently less evidence to substantiate her claims), she might simply repeat similar words and phrases throughout her essay without bringing in outside information. On the other hand, if Josie has high knowledge on this topic, she might be more likely to bring in outside information and only repeat certain key words and phrases throughout the essay.

This example highlights important differences that must be considered when modeling the writing processes engaged by students, which may ultimately contribute to more nuanced assessments of their performance. For example, while surface-level features of a student's essay may be able to be modeled with more traditional, static NLP metrics (e.g., word frequency), the coherence of their writing may require the writer to distribute topical information and outside evidence in specific ways. The distributions of this information may therefore be missed in the absence of dynamical analyses.

Recurrence Quantification Analysis

In the current paper, we use Recurrence Quantification Analysis (RQA) to quantify the extent to which recurrent patterns in students' persuasive essays relate to expert ratings of their quality and characteristics. RQA is a nonlinear technique that provides information about patterns of repetitive behavior in continuous or categorical time series (Marwan, Romano, Thiel, & Kurths, 2007). Similar to many techniques used in dynamical systems theory research, this technique has been used in a variety of domains to characterize temporal patterns of human and non-human behavior (Dale & Spivey, 2005; Marwan, Wessel, Meyerfeldt, Schirdewan, & Kurths, 2002). For instance, RQA has been used to characterize heartrate variability (Marwan et al., 2002), postural fluctuations (Riley, Balasubramaniam, & Turvey, 1999), and eye movements (Anderson et al., 2013).

Recently, researchers have demonstrated that RQA can be applied to categorical data sets and, consequently, be used to provide information about human language (Dale & Spivey, 2005). The fact that this technique can be applied to both continuous and categorical data sets may be particularly important for the study of natural language, because it can measure multiple levels of the text, rather than relying on unidimensional analyses.

RQA analyses begin with the development of a recurrence plot, which is a visualization of a matrix where the individual elements represent points in a time series that are visited more than once. Therefore, the recurrence plot represents the times in which a dynamical system visits the same area in a phase space (Marwan et al., 2007). Each point in the plot represents a particular state that is revisited by the system (e.g., a word). If multiple points occur together, they form *diagonal lines*; these lines represent times when the system revisits an entire sequence of states.

After the recurrence plots are generated, quantitative analyses can be conducted to quantify these plots. RQA calculates numerous indices that quantify recurrent patterns in a particular system (e.g., a text) to allow for statistical comparisons of multiple systems (Zbilut & Webber, 1992; Coco & Dale, 2013 for more information).

Current Study

The current study investigates how and whether information about students' writing performance can be modeled through dynamical analyses of their word use. To this end, we use RQA to calculate seven indices based on the temporal distributions of students' word use. Our aim is to then use these indices to model the holistic quality and characteristics of the essays.

We collected timed, persuasive essays written by undergraduate students and scored by expert human raters. We hypothesized that the RQA indices would provide meaningful information about the writing processes enacted by students, which would subsequently relate to the quality and characteristics of their essays.

Methods

Participants

Undergraduate students (n = 131) from a public university in the United States participated in the study for course credit. On average, the students were 19.8 years in age, with 44.3% identifying as female, 64.1% Caucasian, 14.5% Asian, 7.6% African American, 7.6% Hispanic, and 6.1% "Other."

Data Collection Procedure

Each student wrote a timed (25-minute), persuasive essay in response to a Scholastic Achievement Test (SAT) style prompt. The completed essays contained an average of 412.3 words (*SD* = 159.9, *min* = 47.0, *max* = 980.0).

Essay Scoring

Students' essays were assessed by two independent pairs of expert human raters. These raters had previous experience scoring academic essays and were compensated for their time. The holistic grading rubric was on a 6-point scale and based on a standardized rubric typically used for the assessment of SAT essays. The rubric contained sub-scale scores, which assessed the quality of the following aspects of the essay: *introduction*, *body*, *conclusion*, *word choice*, *sentence structure*, *organization*, *topic and global cohesion*, *voice*, and *grammar, style and mechanics*.

Data Processing

Students' essays were cleaned in preparation for the RQA. All punctuation was removed and the words were converted to lower case and stemmed.

Once the essays were cleaned, the words were converted into series of categorical numeric codes, wherein the codes represented the individual word types (i.e., the unique words) in each essay. For instance, the sentence, "Dogs eat dog food." would be converted to the series: {1, 2, 1, 3}.

Recurrence Quantification Analysis

The *crqa* R library (Coco & Dale, 2013) was used to generate recurrence plots and calculate recurrence indices for the essays. The resulting indices are described in Table 1.

Statistical Analyses

To assess the degree to which the patterns of recurrence in students' essays were associated with their quality, we calculated Pearson correlations and regression analyses between students' essay scores and the RQA indices.

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

Pearson correlations were used to assess relations between word recurrence and essay scores. We calculated these correlations for students' holistic essay scores, as well as the sub-scale scores. Multicollinearity was then assessed among the significantly or marginally significantly correlated indices (r > .90); in the case that two variables demonstrated multicollinearity, the index with the highest correlation with the dependent variable was retained.

A stepwise linear regression analysis was conducted to assess which of the significant RQA indices were most predictive of essay scores. To avoid overfitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed for a maximum of eight indices to be entered in to the model, given that there were 131 essays in the analysis.

Following this essay score analysis, similar follow-up analyses were conducted using the keystroke indices to predict the linguistic features of the essays. For these analyses, we followed the same procedure detailed above.

Table 1. *Description of RQA Indices*

	Description
Recurrence Rate	Density of points in recurrence plot. This metric represents the overall amount of recurrence present in the recurrence plot, regardless of the distributions of the points
Determinism	Number of recurrent points that tend to fall on diagonal lines (ignoring the LOI). This metric provides information about the *distribution* of recurrent points. Systems with low determinism are considered less "ordered" than highly deterministic systems.
Line Number	Number of lines in the recurrence plot. *Lines are defined as two or more consecutive points in a recurrence plot.*
Max Line	Length of the longest diagonal line in the recurrence plot; therefore, this metric reveals if a system revisits a long sequence of states at a particular point in time.
Average Line	Average length of the diagonal lines in the recurrence plot; this metric therefore provides information about the average length of *sequences* of states
Entropy	Shannon entropy of the distribution of the line lengths in the recurrence plot. Entropy will be higher if the system revisits a wider variety of state sequences over time.
Normalized Entropy	Entropy variable normalized by the number of lines in the plot

Results

Recurrence and Essay Quality

Pearson correlations were calculated between the RQA indices and students' holistic essay scores to examine the strength of the relationships among the variables. The results of this analysis identified five RQA indices that demonstrated a significant or marginally significant relation with holistic essay scores (see Table 2).

A linear regression analysis was calculated with these five RQA indices as predictors of students' holistic essay scores (score range: 1-6). This analysis yielded a significant model, R^2 = .432, p < .001, with four variables that combined to account for 43% of the variance in the essay scores: *Log of Line Number* [β = 0.54, p < .001], *Determinism* [β = -.42, p < .001], *Average Line* [β = 0.49, p < .001], and *Max Line* [β = -0.22, p < .05].

These correlation and regression analyses indicate better writers produced essays that contained a higher quantity of recurrent sequences (*Log of Line Number*), as well as longer recurrent sequences, on average (i.e., they had a longer *Average Line Length*, which indicates that lines of recurrent sequences were longer on average). However, these essays were also less deterministic overall, suggesting that these essays contained a higher quantity of individual recurrent points (words) than sequences of words.

Table 2. *Correlations between RQA Indices and Essay Scores*

RQA Index	Holistic	Intro.	Body	Conc.	Word Choice	Sentence Structure	Organization	Cohesion	Voice	Grammar/ Mechanics
Recurrence Rate	-.030	-.101	.005	.031	-.198	-.058	-.022	.064	-.005	-.187
Determinism	-.204	-.006	-.243	-.046	-.214	-.230	-.079	-.152	-.218	-.128
Log of Line Number	.452	.297	.408	.472	.151	.354	.405	.164	.320	.158
Max Line	.145	.224	.173	.096	-.006	.073	.110	-.011	.076	.055
Average Line	.272	.222	.242	.155	.101	.015	.148	.162	.175	.002
Entropy	.224	.189	.139	.209	.032	.110	.175	.118	.113	.197
Normalized Entropy	.014	-.044	-.064	-.015	.001	-.042	-.072	.028	-.006	.159

p <.001 (light gray); p <.05 (medium gray); Marginal (dark gray)

Recurrence and Essay Characteristics

Our second goal was to examine whether the RQA indices were related to the characteristics of the students' essays. Pearson correlations were calculated between the RQA indices and the nine sub-scale essay scores (see Table 2) and followed by regression analyses. The statistical information for these resulting models is provided below.

Introduction Quality

The regression yielded a significant model, R^2 = .117, p < .001 with two significant predictors: *Log of Line Number* [β = 0.27, p < .001] and *Average Line* [β = 0.17, p < .001].

Body Quality

The regression yielded a significant model, R^2 = .384, p < .001 with three significant predictors: *Log of Line Number* [β = 0.43, p < .001], *Determinism* [β = -0.49, p < .001], and *Average Line* [β = 0.38, p < .001].

Conclusion Quality

The regression yielded a significant model, R^2 = .223, p < .001 with one significant predictor: *Log of Line Number* [β = 0.43, p < .001].

Word Choice

The regression yielded a significant model, R^2 = .114, p < .001 with three significant predictors: *Determinism* [β = - 0.20, p < .05], *Log of Line Number* [β = 0.23, p < .05], and *Recurrence Rate* [β = -0.19, p < .05].

Sentence Structure

The regression yielded a significant model, R^2 = .215, p < .001 with two significant predictors: *Log of Line Number* [β = 0.41, p < .001] and *Determinism* [β = -0.30, p < .001].

Organization

The regression yielded a significant model, R^2 = .164, p < .001 with one significant predictor: *Log of Line Number* [β = 0.41, p < .001].

Topic and Global Cohesion

None of the RQA indices were entered into this analysis.

Voice

The regression yielded a significant model, R^2 = .250, p < .001 with three significant predictors: *Log of Line Number* [β = 0.34, p < .001], *Determinism* [β = -0.41, p < .001], and *Average Line* [β = 0.29, p < .01].

Grammar, Style, and Mechanics

The regression yielded a significant model, $R^2 = .106$, $p < .01$ with two significant predictors: *Entropy* [$\beta = 0.28$, $p < .01$], and *Recurrence Rate* [$\beta = -0.27$, $p < .01$].

The results of the sub-scale analyses indicate that the RQA indices were meaningfully related to the properties of students' essays at multiple levels; yet, the sub-scale scores were more weakly related than the holistic essay scores. The regression analysis calculated for *body quality* was the strongest of the models and indicated that essays with higher-quality body paragraphs were related to longer lines, but lower determinism overall. Additionally, the regressions for all remaining sub-scales, except cohesion, were significant with RQA indices accounting for between 11 and 25% of the variance. The *topic and global cohesion* score was not significantly related to any of the RQA indices, indicating that human perceptions of cohesion were not related to recurrent word patterns in students' writing.

Discussion

In the current study, we used dynamic methodologies to develop NLP assessments of students' writing performance. In particular, our goal was to determine whether we could model the holistic and sub-scale scores of essays by calculating indices related to the temporal distribution of the words that they produced. Recurrence quantification analysis (RQA) was used to calculate indices related to the quantity, length, and distributions of these recurrent word patterns. The results revealed that the RQA indices were able to model 43% of the variance in students' holistic essay scores. Additionally, these indices were able to model specific characteristics of the essays at multiple levels.

The essay score analyses revealed that five RQA indices were significantly or marginally significantly correlated with students' holistic essay scores. This finding is promising and indicates that the overall quality of students' essays can be modeled in the ways in which they distribute the words throughout their writing. These initial analyses of essay score indicate that the length and variability of the word sequences that students produce may be an important indicator of their writing skill. In particular, higher-quality essays were characterized by longer word sequences, but also by a greater variability in these line lengths (Entropy) and lower Determinism overall. These analyses speak to the importance of accounting for temporal patterns in language. NLP techniques often rely on summative metrics of text features to characterize student writing; however, these results suggest that expanding these analyses to include temporal information can provide insights into the characteristics of quality writing.

The essay characteristic analyses additionally revealed similarities and differences between the RQA indices and the quality of students' writing at multiple levels. Performance on all of the essay scores was related to a greater number of recurrent word sequences (Log of Line Number), and all of the paragraph quality metrics demonstrated significant correlations with *Average Line Length* and *Maximum Line Length.* This finding suggests that writing quality at multiple levels can be characterized by a greater amount of recurrent information throughout the text, as well as longer sequences of these recurrent words.

Beyond these similarities, the correlations were indicative of differences among the essay scores. Specifically, while the relations between the RQA indices and the essay scores were generally similar in their directionality, they largely differed in magnitude. For instance, the quality of students' body paragraphs was significantly related to four of the seven RQA indices, whereas the topic and global cohesion score demonstrated only three marginally significant correlations with the indices. These results suggest that these recurrent word patterns can provide fine-grained information about writing quality that moves beyond holistic scores.

The results of the current study provide initial evidence for the usefulness of dynamic analyses of language. However, there remain a number of open questions to be answered in additional research. For instance, do these indices map onto similar quality metrics across multiple text genres? Similarly, do the indices predict writing quality for different age groups and native languages? These questions and many more remain to be answered in future research.

An important note is that our analyses only focused on the individual words in students' essays. We did not account for the various additional sources of information that are currently afforded in NLP analyses, such as parts-of-speech, semantic information, or sophistication. We strategically chose to focus this initial study on the individual words in order to provide a demonstration of the strength of RQA in the absence of this additional information. However, this study by no means represents the limit of its potential. RQA is a highly flexible technique that can be used to analyze any temporal data -- continuous or categorical. For instance, one could imagine examining recurrent patterns in the topics discussed in students' essays, the parts-of-speech or the sophistication of the words. Future analyses such as these will no doubt provide important insights into the structure of student language.

Overall, our results suggest that RQA can be utilized to guide dynamic assessments of students' writing quality. Our eventual goal is to use these indices to develop more nuanced assessments of essay quality, which can then be used to drive formative feedback in adaptive educational technologies. Although this study provides only a first step toward that goal, and a number of future research remains to be conducted, these results provide a foundation on which to conduct research that considers the dynamic nature of student language.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120707 to Arizona State University. The opinions expressed are those of the authors and do not necessarily reflect the views of the IES.

References

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., and McNamara, D. S. 2016. Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. P. Rosé (Eds.), *Proceedings of the 6th International Learning Analytics & Knowledge Conference, Edinburgh, United Kingdom (LAK'16)*, (pp. 114-123). New York, NY: ACM.

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)*, (pp. 246-254). Poughkeepsie, NY: ACM.

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), pp.842-856.

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Brysbaert, M., Warriner, A. B. and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46, 904–911.

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*.

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

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* (pp. 530-535). Mahwah, NJ: Lawrence Erlbaum.

Dikli, S. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*.

D'Mello, S., Dowell, N., 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.* (pp. 9–16). IOS Press, Amsterdam.

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.

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

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.

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* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. 2015. Hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35-59.

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

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

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

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

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.

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.