MODELING INDIVIDUAL DIFFERENCES AMONG WRITERS USING READERBENCH

MODELING INDIVIDUAL DIFFERENCES AMONG WRITERS USING READERBENCH

Laura K. Allen[1], Mihai Dascalu[2], Danielle S. McNamara[1], Scott A. Crossley[3], Stefan Trausan-Matu[2]

[1] *Arizona State University (UNITED STATES)*
[2] *University Politehnica of Bucharest (ROMANIA)*
[3] *Georgia State University (UNITED STATES)*

Abstract

The current study builds upon a previous study, which examined the degree to which the lexical properties of students' essays could predict their vocabulary scores. We expand on this previous research by incorporating new natural language processing indices related to both the surface- and discourse-levels of students' essays. Additionally, we investigate the degree to which these NLP indices can be used to account for variance in students' reading comprehension skills. We calculated linguistic essay features using our framework, *ReaderBench*, which is an automated text analysis tools that calculates indices related to linguistic and rhetorical features of text. University students (n = 108) produced timed (25 minutes), argumentative essays, which were then analyzed by *ReaderBench*. Additionally, they completed the Gates-MacGinitie Vocabulary and Reading Comprehension tests. The results of this study indicated that two indices were able to account for 32.4% of the variance in vocabulary scores and 31.6% of the variance in reading comprehension scores. Follow-up analyses revealed that these models further improved when only considering essays that contained multiple paragraph (R^2 values = .61 and .49, respectively). Overall, the results of the current study suggest that natural language processing techniques can help to inform models of individual differences among student writers.

Keywords: Writing skill, automated writing evaluation, comprehension prediction, vocabulary measures, natural language processing.

1 INTRODUCTION

Effective written communication is a complex cognitive and social skill that is important for success in educational and work-place settings [1]. To generate text that successfully communicates a particular idea, a writer must have knowledge of the language and topic, and also know how to flexibly adapt their language to suit different audience and contexts [2; 3]. It may come as no surprise, then, that teaching students to become strong writers is not a simple task. Evidence for this fact comes from national and international assessments that consistently highlight students' struggles on writing tasks [1; 4; 5].

To effectively foster strong writing skills, teachers need to provide explicit instruction to their students that describes and demonstrates the skills and strategies that are needed during the prewriting, drafting, and revision phases of the writing process. Further, this instruction should be accompanied by ample opportunities for students to engage in *deliberate practice*, through the delivery of both summative and formative essay feedback. Deliberate practice is a critical component of this a successful writing curriculum [6; 7], because it provides students with actionable suggestions for revising their writing, and can help to promote better regulation of the phases of the writing process [7]. One significant problem with these instructional goals, however, is that they rely on teachers to frequently generate individualized feedback on their students' essays. This task can be extremely difficult due to large class sizes and limited time to provide detailed comments every time students are asked to write an essay.

To address these complex goals and challenging demands, many educators and researchers are increasingly exploring the use of computer-based instructional systems that can provide automated assessments of the quality and characteristics of students' writing and deliver meaningful feedback in the absence of a teacher [8]. Specifically, automated essay scoring (AES) systems calculate a multitude of linguistic properties of students' essays and use these to assign holistic scores [9; 10]. These systems use a multitude of natural language processing (NLP) and machine learning methodologies to provide these essay scores, and previous research suggests that they are often

comparable to human raters [10; 11; 12; 13]. Within educational contexts, AES systems are commonly incorporated into educational learning environments, such as automated writing evaluation (AWE) systems [14] and intelligent tutoring systems (ITSs) [15] to provide students with feedback and instruction related to these scores.

It is important to note that these writing evaluation systems are primarily focused on modeling information about the *quality* of individual essays that students submit. While this is obviously an important goal, this limited focus may have negative consequences on the potential efficacy of the instruction delivered by these systems. Specifically, while prior research has suggested that these scores are highly valid and reliable [10; 11; 12; 13] critics suggest that these accurate scores will only go so far in helping students to learn to improve their writing. Instead, they suggest that computer-based writing instruction will need to become more personalized to have a significant impact on educational outcomes.

One approach to increasing system personalization is to incorporate information about students that can be used to guide more beneficial feedback. For example, previous studies have shown that *vocabulary knowledge* plays a major role in the writing process because it is strongly correlated with the scores assigned to students' essays [2; 16]. Writers with a strong knowledge of vocabulary will likely have an easier time generating lexically sophisticated essays than students who struggle in this area. Given this fact, it may be the case that students with high and low vocabulary knowledge should be given different forms of automated essay feedback, even if they receive the same essay scores. The holistic quality of the essays may be similar across these two students; however, the factors contributing to this quality and the steps needed to improve the quality may largely differ. Therefore, feedback delivery may be more effective if it takes student-level information into account, in addition to essay scores.

Expanding research beyond the modeling of essay scores to the modeling of individual differences could be an important step. Specifically, these models could help to drive more personalized instruction and feedback that is tailored to students' strengths and weaknesses. In the current paper, we examine the efficacy of NLP techniques to inform stealth assessments of individual differences. In particular, we examine whether the linguistic properties of students' essays can accurately model their scores on standardized measures of vocabulary knowledge and comprehension skill. Our ultimate goal is to use these measures to provide more individualized tutoring to student users.

1.1 Stealth Assessments and Adaptive Instruction

Computer-based learning environments often rely on repeated assessments of students' performance in order to increase the adaptivity of their instruction and feedback. These measures are important because they deliver detailed information about students, such as their knowledge states, motivation levels, cognitive processes, which can help systems decide when and how to provide specific forms of remediation or recommendations. One problem, however, is that these repeated assessments can often be harmful to students' learning. This constant exposure to surveys and tests can have a significant, negative impact on "flow" [17] and, consequently, decrease subsequent performance.

To address this issue, researchers have attempted to develop novel methods for accumulating information about individual student users without causing disturbances for the learning experience [17; 18]. Specifically, some researchers have highlighted the importance of developing "stealth assessments," which are assessments that have been embedded in the learning task and are not able to be explicitly detected by the students [19] These stealth assessments can be informed by a wealth of information that can be easily logged in the system, such as the speed at which someone is typing to the trajectories of their mouse movements [8; 20]. Once developed, these stealth assessments can be added to student models as a means to provide more individualized instruction and feedback [21].

Natural language processing (NLP) is one methodological technique that may provide critical information to inform these stealth assessments. NLP involves the automated calculation of linguistic indices of texts using a computer program (or programming language) [22]. Thus, the focus of NLP is largely centered around the use of computers to understand and produce natural language in order to automate certain tasks (e.g., chat bots that provide help on certain websites) or study communicative processes (e.g., to investigate the linguistic properties of high-quality tests).

Importantly, a number of NLP tools have been developed to calculate indices at multiple levels of text (e.g., lexical, syntactic, discourse), which may be particularly beneficial for researchers who are interested in examining the writing process [23] or for many other domains in which students produce

natural language [24]. Previous research with NLP has provided critical information about the learning process, across a number of domains and educational contexts. With respect to the study of writing, NLP has been particularly informative, primarily as a means of modeling human ratings of text quality [12; 23; 25].

1.2 *ReaderBench*, an Integrated Discourse Analysis Framework

ReaderBench [26] is a fully functional automated software framework, designed to provide support for students and tutors in terms of comprehension assessment and prediction in various educational scenarios. The system makes uses of text-mining techniques based on advanced natural language processing and machine learning algorithms to design and deliver summative and formative assessments using multiple data sets (e.g., textual materials, behavior tracks, self-explanations).

ReaderBench targets both tutors and students by providing an integrated learning model approach including individual and collaborative learning methods, cohesion-based discourse analysis [27], dialogical discourse models [28], textual complexity evaluation [29], reading strategies identification [30], and participation and collaboration assessment [31]. By using natural language processing techniques, the main purpose of this framework is to bind traditional learning methods with new trends and technologies to support computer supported collaborative learning (CSCL). *ReaderBench*, by design, is not meant to replace the tutor, but to scaffold both tutors and learners by enabling continuous assessment, self-assessment, collaborative evaluation of individuals' contributions, as well as the analysis of reading materials to match readers to an appropriate level of text difficulty.

From a learner's perspective, *ReaderBench* can act as a Personal Learning Environment (PLE) that incorporates: a) *individual assessment* of textual materials making use of the textual complexity metrics (semantics, morphology, surface factors integrated by support vector machines) that reflect the textual organization and structure of reading materials [29]; b) *comprehension prediction* by identifying reading strategies employed by students in their self-explanations or by automatically evaluating student summaries [30]; c) *collaboration* and *participation evaluation* in CSCL conversations based on cohesion graphs and on Bakhtin's dialogism [31]. In this paper we will focus on the use of the textual complexity metrics implemented in *ReaderBench* to predict essay quality.

2 CURRENT STUDY

The current study builds upon a previous study, which provided an initial examination into our research goals [32]. The purpose of this prior study was to investigate the degree to which the lexical (word-level) features of students' essays could account for variance in their performance on a vocabulary knowledge test. Indices related to these lexical properties were calculated with TAALES, which is an automated text analysis tool that provides information about the lexical sophistication of texts. Results of this initial study revealed that two lexical indices were able to account for 44% of the variance in college students' vocabulary knowledge scores.

The purpose of the current study is to expand on this previous analysis by incorporating new natural language processing indices related to both the surface- and discourse-levels of students' essays. We further extend this previous study by including an additional dependent variable, comprehension skill, to examine whether and how the indices perform on a different performance metric.

To accomplish this goal, the linguistic and rhetorical properties of the students' essays from the previous study were calculated using the *ReaderBench* framework. We hypothesized that the lexical properties of the essays would provide significant predictive power in modeling both students' performance on the vocabulary and comprehension tests, similar to the previous analysis. Additionally, we hypothesized that the discourse-level indices calculated by *ReaderBench* would account for additional variance in addition to these surface-level features.

2.1 Corpus

The current corpus consisted of 108 essays written by undergraduate students from a large university campus in the Southwest United States. On average, these students were 19.75 years of age (range: 18-37 years), and the majority of these students identified as freshmen or sophomores. Of the 108 students, 52.9% were male, 53.7% were Caucasian, 22.2% were Hispanic, 10.2% were Asian, 3.7% were African-American, and 9.3 % reported other ethnicities. All students wrote a 25-minute prompt-based, persuasive essay that resembled the demands of standardized testing settings. The students

were not allowed to proceed until the entire 25 minutes had elapsed. The essays in the corpus contained an average of 410.44 words (*SD* = 152.50), ranging from a minimum of 84 words to a maximum of 984 words.

2.2 Vocabulary Knowledge Assessment

The vocabulary knowledge of the students was assessed using the vocabulary portion of the Gates-MacGinitie (4th ed.) reading test (form S) level 10/12 [1]. This assessment is a 10-minute task, which is comprised of 45 simple sentences that each contains an underlined vocabulary word. Students were asked to read each sentence and then select the most closely related word (from a list of five choices) to the underlined word within the sentence.

2.3 Comprehension Skill Assessment

Students' reading comprehension skills were measured using reading comprehension section of the Gates-MacGinitie (4th ed.) reading test (form S) level 10/12 [1]. 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

To assess the properties of students' essays, we opted to use developed and validated textual complexity model from the *ReaderBench* framework [27; 29], which integrates a multitude of indices ranging from classic readability formulas, surface indices, morphology and syntax, as well as semantics. A unique contribution of *ReaderBench* is that is explicitly focuses on text cohesion and discourse connectivity, and provides a more in-depth analysis of the structure of discourse based on Cohesion Network Analysis (CNA) [33]. The purpose of CNA is to model the semantic links between different text constituents in a multi-layered cohesion graph (similar to Social Network Analysis) [34]. Additionally, *ReaderBench* computes *Age of Exposure* [35], which is a novel word complexity metric that approximates the typical learning curve of individual words based on incremental (additive) Latent Dirichlet Allocation (LDA) models [36]. Intermediary spaces are automatically aligned to the most mature LDA model in order to assess the degree to which a concept has been properly understood. Below, we provide a more thorough description of the *ReaderBench* indices that were chosen to be included in the current study. See [37] for more through explanations of each variable.

Surface-level text indices. The first automated measures of text complexity were developed by Page [38] in his search to develop an automatic grading system for students' essays. Page discovered a strong correlation between human intrinsic variables (trins) and proxies (i.e., the text complexity indices), thus providing evidence that statistical analyses can provide reliable estimations of important textual features. Our model integrates the most representative and predictive proxies from Page's initial study, along with additional surface-level measures that are frequently used in other automated essay grading systems (e.g., average word/phrase/contribution length, average unique/content words per contribution, average commas per sentence/contribution). For example, *word entropy*, derived from Shannon's Information Theory [39], is a relevant metric for quantifying textual complexity based on the hypothesis that a more complex text contains more information and more diverse concepts. In contrast, character entropy is a language specific characteristic [40] and does not typically account for a significant amount of variance in English texts.

Word complexity. *ReaderBench* calculates a number of indices that represent different layers of discourse analysis to estimate the difficulty of individual words: a) syllable count, b) distance in characters between the inflected form, lemma and word stem (adding multiple prefixes or suffixes increases the difficulty of using a certain word), c) inverse document frequency from a text corpus (in our case, the TASA corpus - http://lsa.colorado.edu/spaces.html), d) the distance in the hypernym tree and e) the word polysemy count from WordNet [41].

Syntactic complexity indices. In order to quantify textual complexity in terms of syntactic structure, *ReaderBench* calculates multiple indices related to the different parts of speech (POS). Discourse is most predominantly analyzed in terms of nouns and verbs only; however, our framework aims to extend this previous research by incorporation information about the prepositions, adjectives and adverbs, because these POS can provide important information about the complex structure of the discourse. According to Gervasi and Ambriola [40], sentences that contain higher numbers of semantic dependencies, or that have more depth in their parsing trees are indicative of more complex

discourse structures; thus, additional indices extracted from sentence parsing trees (e.g., the *number of nodes, dependencies or the maximum tree depth*) have been integrated into *ReaderBench*.

Semantic cohesion indices. In order to comprehend texts, readers must create coherent and well-connected representations of the information it contains. This mental representation is commonly referred to as the situation model [42]. According to McNamara et. al. [43], the cohesion of a text plays an important role in the ease with which readers create coherent text representations, and therefore, comprehend the text. Therefore, *ReaderBench* places an emphasis on the calculation of multiple indices of text cohesion. Specifically, both local and global indices of cohesion are computed based on the CNA graph of the text, based on the semantic similarities of the links [34] at both intra- and inter-paragraph levels. Cohesion is estimated as the average value of [27]: a) Wu-Palmer semantic distances applied on the *WordNet* lexicalized ontology, b) cosine similarity in Latent Semantic Analysis (LSA) vector space models, and c) the inverse of the Jensen Shannon dissimilarity (JSD) [44] between Latent Dirichlet Allocation (LDA) topic distributions.

2.5 Statistical Analyses

Statistical analyses were conducted to investigate the ability of *ReaderBench* indices to assess and model students' scores on assessments of their vocabulary knowledge and text comprehension skills. Pearson correlations were first separately calculated between the linguistic properties of students' essays (as assessed by *ReaderBench*) and their scores on vocabulary and comprehension tests. For each analysis, the indices that demonstrated a significant or marginally significant correlation with the assessment score (i.e., vocabulary or comprehension) were retained in the analysis. Multicollinearity was then assessed among the indices ($r > .90$). When two or more indices demonstrated multicollinearity, the index that correlated most strongly with the relevant assessment score was retained in the analysis. All remaining indices were finally checked to ensure that they were normally distributed.

Two stepwise regression analyses were separately conducted to assess which of the remaining linguistic indices were most predictive of vocabulary knowledge and comprehension skills. To avoid overfitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed 7 indices to be entered, given that there were 108 essays included in the analysis. 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 increase the probability that the results could be generalized to a new data set.

3 RESULTS

3.1 Vocabulary Knowledge Analyses

Pearson correlations were calculated between the *ReaderBench* indices and students' Gates-MacGinitie vocabulary knowledge scores to examine the strength of the relationships among these variables. The analysis indicated that there were 11 indices that demonstrated a significant correlation with vocabulary knowledge scores. We selected the 7 indices that were most strongly correlated with performance on the vocabulary test and did not demonstrate multicollinearity with each other.

Table 1. Correlations between Gates-MacGinitie vocabulary knowledge scores and *ReaderBench* linguistic indices.

ReaderBench variable	*r*	*p*
Average Word Length	0.510	<.001
Average Syllables per Word	0.483	<.001
Average Difference in Characters between Inflected Word Forms and Corresponding Stems	0.460	<.001
Average Polysemy Senses of Each Word	-0.423	<.001
Word Entropy	0.374	<.001
Average Number of Pronouns per Paragraph	-0.192	<.05
Average Number of Adjectives per Sentence	0.186	=.05

Table 1 presents the Pearson correlations between scores on the vocabulary knowledge test and the 7 linguistic indices selected for the regression analysis. Vocabulary scores were primarily related to the lexical properties of students' essays, such as the average length and number of syllables per word. Additionally, vocabulary scores were related to the diversity of words that writers include in the essays, as evidenced by the *word entropy* and *average number of words with different word stems* indices.

A stepwise regression analysis using the 7 indices listed in Table 1 as the predictors of vocabulary knowledge scores yielded a significant model for the students in the training set, $F(2, 76) = 18.243$, $p < .001$, $r = .570$, $R^2 = .324$. Two of the 7 indices were significant predictors in the regression analysis and combined to account for 32% of the variance in students' vocabulary scores: average word length [$\beta = .50$, $t(76) = 5.256$, $p < .001$] and word entropy [$\beta = .27$, $t(76) = 2.822$, $p < .01$]. The regression model for the training set is presented in Table 2. The test set yielded $r = .747$, $R^2 = .558$, resulting in a better fit to the data than the training model and accounted for 56% of the variance in vocabulary knowledge scores.

Table 2. *ReaderBench* regression analysis predicting Gates-MacGinitie vocabulary knowledge scores.

Entry	Variable Added	R^2	ΔR^2
Entry 1	Average Word Length	0.254	0.254
Entry 2	Word Entropy	0.324	0.071

Overall, the results of this regression analysis indicate that the students who performed better on the vocabulary knowledge assessment generated essays that were more lexically sophisticated. The essays contained words that were longer and also had a higher word entropy. Therefore, writers with stronger vocabulary knowledge tended to use longer words with a higher diversity (entropy) throughout their essays.

3.2 Comprehension Skill Analyses

Our second analysis examined the degree to which the *ReaderBench* indices could accurately model students' scores on the comprehension assessment. Pearson correlations were calculated between the *ReaderBench* indices and students' Gates-MacGinitie reading comprehension scores. Eleven indices were significantly correlated. As with the vocabulary analysis, we selected the 7 indices that were most strongly correlated with comprehension and did not exhibit multicollinearity. These indices are listed in Table 3.

Table 3. Correlations between Gates-MacGinitie reading comprehension scores and *ReaderBench* linguistic indices.

ReaderBench variable	*r*	*p*
Average Word Length	0.508	<.001
Average Syllables per Word	0.485	<.001
Average Difference in Characters between Inflected Word Forms and Corresponding Stems	0.465	<.001
Average Polysemy Senses of Each Word	-0.387	<.001
Word Entropy	0.335	<.001
Average Number of Adjectives per Sentence	0.280	<.01
Average Number of Commas per Sentence	0.226	<.05

A stepwise regression analysis was calculated for the students in the training set with the indices listed in Table 3 as the predictors of students' comprehension scores. This regression resulted in a significant model, $F(2, 76) = 17.533$, $p < .001$, $r = .562$, $R^2 = .316$ with two significant predictors: *average word length* [$\beta = .51$, $t(76) = 5.395$, $p < .001$] and *word entropy* [$\beta = .22$, $t(76) = 2.284$, $p < .05$]. This model is presented in Table 3. The test set yielded $r = .727$, $R^2 = .528$, resulting in a better fit to the data than the training model, similar to the vocabulary knowledge assessment.

Table 4. *ReaderBench* regression analysis predicting Gates-MacGinitie reading comprehension scores.

Entry	Variable Added	R^2	ΔR^2
Entry 1	Average Word Length	0.269	0.269
Entry 2	Word Entropy	0.316	0.047

The results of the comprehension analyses revealed strong similarities with the vocabulary analyses. In particular, students' performance on the comprehension assessment was positively correlated with, and best modeled by, indices related to word length and diversity (word entropy). Therefore, more skilled text comprehenders produced essays that were more lexically sophisticated than less skilled comprehenders.

3.3 Follow-up Analyses

In the previous analyses, we were able to confirm our first hypothesis; however, we failed to confirm our second. Namely, the results revealed that both vocabulary and comprehension scores were best modeled by surface-level properties of students' essays, rather than discourse-level indices. Obviously, one potential explanation for this finding is that discourse-level indices do not carry predictive information about these individual differences. An alternative explanation, however, relates to the scope of the discourse-level indices. In the previous analyses, we focused on a complete corpus of student essays that contained both one-paragraph and multiple-paragraph essays. This approach significantly limited our ability to rely on many of the discourse-level indices available in *ReaderBench.* Because they aim to capture information about essays a broad level, many of the discourse-level indices rely on the assumption that essays contain multiple paragraphs that can be analyzed. For example, *ReaderBench* provides an index of semantic similarity among the introduction, body, and conclusion paragraphs of an essay.

We next conducted follow-up exploratory analysis as a preliminary test of this explanation. Specifically, we conducted similar regression analyses for a subset of our corpus that contained essays with 3 or more paragraphs (n = 57). This approached allowed us to conduct an investigation into the potential usefulness of these semantic cohesion indices for multiple-paragraph texts.

As with the previous analyses, Pearson correlations were separately calculated between the *ReaderBench* indices and students' Gates-MacGinitie vocabulary knowledge and reading comprehension scores. For each dependent variable, the 7 indices that demonstrated the strongest correlations (and no multicollinearity) were retained as predictors in the regression analyses.

The results of the vocabulary regression analysis indicated that four indices were able to account for 61% of the variance in vocabulary scores [$F(4, 52) = 19.985$, $p < .001$; $R^2 = .606$]: average polysemy senses of each word, word entropy, average word length, and semantic similarity among text segments, namely the first paragraph and the middle section(s) measured by LSA. Similarly, average word length, word entropy, and semantic similarity among text segments (similarly start-middle configuration) accounted for 49% of the variance in comprehension scores [$F(3, 53) = 16.908$, $p < .001$; $R^2 = .489$]. These analyses provide preliminary evidence of our hypothesis that for essays with 3 or more paragraphs, semantic cohesion indices were able to account for unique variance in our dependent variables and, consequently, increase the predictive power of our models.

4 DISCUSSION

Our aim in the current study was to build upon a previous research study to that aimed to assess and model individual differences among students by calculating the lexical properties of their essays. We employed *ReaderBench,* an automated text analysis tool to calculate linguistic indices for the essays that extended beyond lexical features (e.g., syntactic complexity, semantic cohesion). Correlation and regression analyses were then conducted with these indices to determine whether they could provide predictive power to drive stealth assessments of vocabulary knowledge and comprehension skills.

The initial correlation analyses identified a number of *ReaderBench* indices that were significantly correlated with students' vocabulary knowledge and comprehension skills. Further, the regression analyses revealed that indices related to the length and diversity of the words in the essays combined to account for between 32% and 56% of the variance in these scores. These analyses generated

similar results as [32], and suggest that students with stronger vocabulary knowledge and comprehension skills tended to generate essays that contained longer words. Additionally, these students were more likely to include a greater variety of words in their writing.

The follow-up analyses provided additional information about additional indices that may carry provide an important role in the development of stealth assessments. Vocabulary knowledge and comprehension skill were better characterized by our models when we only considered essays that contained three or more paragraphs. This result is largely due to the fact that we were able to consider a larger number of *ReaderBench* indices that tap into discourse level information. These results are preliminary, however, and require further research to be confirmed. Future research should specifically test these assumptions and consider developing separate stealth assessments for single and multiple-paragraph essays.

An additional goal of future research is to consider further individual difference variables that may be more difficult to test. Here, we focused on vocabulary knowledge and comprehension skill because these scores have been shown to be strongly related to writing proficiency [2]. In future studies, however, it will be important to consider other individual differences, and to determine the generalizability of our methodology. For instance, we might choose to model a wide variety of factors related to writing performance, such as students' attitudes and self-efficacy towards writing, their motivation level on a particular day, or their level creativity. Some of these indices may be relatively simple to assess using the NLP techniques outlined in this paper; others, however, may not reliably relate to the properties of students' writing. Instead, more variable factors, such as daily motivation, may be better captured by analyses that focus on *changes* in students' writing (i.e., a comparison of their writing on a particular day to their style and quality of writing more generally), rather than on properties of individual texts.

In conclusion, the current study utilized the NLP framework *ReaderBench* to investigate the efficacy of NLP techniques to inform stealth assessments of vocabulary knowledge. Eventually, we expect that these stealth assessments will enhance student models within computer-based writing instruction systems and allow researchers and educators to provide students with more pointed feedback and instruction. More broadly, these results provide evidence that NLP techniques can be used to help researchers and system developers build stealth assessments. These assessments could be used to inform student models, which will guide the delivery of more personalized instruction and feedback for student users.

ACKNOWLEDGEMENTS

The research reported here was partially supported by FP7 208-212578 LTfLL project, by the 644187 RAGE H2020-ICT-2014, as well as by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080589 to Arizona State University.

REFERENCES

[1] National Commission on Writing. (2003). *The Neglected "R."*. New York, NY: College Entrance Examination Board.

[2] Allen, L.K., Snow, E.L., Crossley, S.A., Jackson, G.T., & McNamara, D.S. (2014). Reading comprehension components and their relation to the writing process. *L'année psychologique/Topics in Cognitive Psychology, 114*, 663–691.

[3] Allen, L.K., Snow, E.L., & McNamara, D.S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis & B. M. McLaren (Eds.), *7th Int. Conf. on Educational Data Mining* (pp. 304–307). London, UK.

[4] National Assessment of Educational Progress. (2011). The Nation's Report Card: Writing 2011. Washington, DC: National Assessment of Educational Progress.

[5] Baer, J. D., & McGrath, D. (2007). The reading literacy of U.S. fourth-grade students in an international context: Results from the 2001 and 2006 Progress in International Literacy Study (PIRLS). Washington, DC: National Center for Educational Statistics, Institute of Education Sciences, U.S. Department of Education.

[6] Johnstone, K.M., Ashbaugh, H., & Warfield, T.D. (2002). Effects of repeated practice and contextual writing experiences on college students' writing skills. *Journal of Educational Psychology, 94*, 305–315.

[7] Kellogg, R., & Raulerson, B. (2007). Improving the writing skills of college students. *Psychonomic Bulletin and Review, 14*, 237–242.

[8] Allen, L.K., Snow, E.L., & McNamara, D.S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* (pp. 246–254). Poughkeepsie, NY: ACM.

[9] Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

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

[11] Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1).

[12] McNamara, D.S., Crossley, S.A., Roscoe, R., Allen, L.K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59.

[13] Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10*, 1–24.

[14] Crossley, S.A., Varner, L.K., Roscoe, R.D., & McNamara, D.S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)* (pp. 269–278). Memphis, USA: Springer.

[15] Roscoe, R.D., Varner, L.K., Weston, J.L., Crossley, S.A., & McNamara, D.S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59.

[16] Crossley, S.A., Allen, L.K., Snow, E.L., & McNamara, D.S. (2015). Pssst…textual Features… there is more to automatic essay scoring than just you! In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* (pp. 203–207). Poughkeepsie, NY: ACM.

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

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

[19] Shute, V.J., Ventura, M., Bauer, M.I., & 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* (pp. 295–321). Mahwah, NJ: Routledge.

[20] Snow, E.L., Jacovina, M.E., Allen, L.K., Dai, J., & McNamara, D.S. (2014). Entropy: A stealth assessment of agency in learning environments. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.), *7th Int. Conf. on Educational Data Mining,* (pp. 241–244). London, UK.

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

[22] Crossley, S.A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46*, 256–271.

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

[24] Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

[25] Varner, L.K., Roscoe, R.D., & 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*, 35–59.

[26] Dascalu, M., Stavarache, L.L., Dessus, P., Trausan-Matu, S., McNamara, D.S., & Bianco, M. (2015). ReaderBench: The Learning Companion. In *17th Int. Conf. on Artificial Intelligence in Education (AIED 2015)* (pp. 915–916). Madrid, Spain: Springer.

[27] Dascalu, M. (2014). Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence (Vol. 534). Cham, Switzerland: Springer.

[28] Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D.S. (2015). Dialogism: A Framework for CSCL and a Signature of Collaboration. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine & S. Ludvigsen (Eds.), *11th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2015)* (pp. 86–93). Gothenburg, Sweden: ISLS.

[29] Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 335–377). Cham, Switzerland: Springer.

[30] Dascalu, M., Dessus, P., Bianco, M., & Trausan-Matu, S. (2014). Are Automatically Identified Reading Strategies Reliable Predictors of Comprehension? In S. Trausan-Matu, K. E. Boyer, M. Crosby & K. Panourgia (Eds.), *12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014)* (pp. 456–465). Honolulu, USA: Springer.

[31] Dascalu, M., Trausan-Matu, S., & Dessus, P. (2014). Validating the Automated Assessment of Participation and of Collaboration in Chat Conversations. In S. Trausan-Matu, K. E. Boyer, M. Crosby & K. Panourgia (Eds.), *12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014)* (pp. 230–235). Honolulu, USA: Springer.

[32] Allen, L.K., & McNamara, D.S. (2015). You are your words: Modeling students' vocabulary knowledge with natural language processing. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura & M. Desmarais (Eds.), *8th Int. Conf. on Educational Data Mining* (pp. 258–265). Madrid, Spain.

[33] Dascalu, M., Trausan-Matu, S., McNamara, D.S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423. doi: 10.1007/s11412-015-9226-y

[34] Trausan-Matu, S., Dascalu, M., & Dessus, P. (2012). Textual complexity and discourse structure in Computer-Supported Collaborative Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis & K. Panourgia (Eds.), *11th Int. Conf. on Intelligent Tutoring Systems (ITS 2012)* (pp. 352–357). Chania, Grece: Springer.

[35] Dascalu, M., McNamara, D.S., Crossley, S.A., & Trausan-Matu, S. (2016). Age of Exposure: A Model of Word Learning. In *30th AAAI Conference on Artificial Intelligence* (pp. 2928–2934). Phoenix, AZ: AAAI Press.

[36] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

[37] Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. In *23rd Int. Conf. on Computational Linguistics (COLING 2010)* (pp. 276–284). Beijing, China: ACL.

[38] Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10*(2), 45–58.

[39] Shannon, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal, 30*, 50–64.

[40] Gervasi, V., & Ambriola, V. (2002). Quantitative assessment of textual complexity. In M. L. Barbaresi (Ed.), *Complexity in language and text* (pp. 197–228). Pisa, Italy: Plus.

[41] Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

[42] van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

[43] McNamara, D.S., Graesser, A.C., & Louwerse, M.M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R&L Education.

[44] Manning, C.D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.