Pssst… Textual Features… There is More to Automatic Essay Scoring than Just You!

Scott Crossley
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
01+404-413-5179
scrossley@gsu.edu

Laura K. Allen
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
LauraKAllen@asu.edu

Erica L. Snow
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
Erica.L.Snow@asu.edu

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
01+404-413-5200
dsmcnamara1@gmail.com

ABSTRACT
This study investigates a new approach to automatically assessing essay quality that combines traditional approaches based on assessing textual features with new approaches that measure student attributes such as demographic information, standardized test scores, and survey results. The results demonstrate that combining both text features and student attributes leads to essay scoring models that are on par with state-of-the-art scoring models. Such findings expand our knowledge of textual and non-textual features that are predictive of writing success.

Categories and Subject Descriptors
K.3.1 [**Computer Uses in Education**]: Computer-assisted Instruction (CAI); J.5 [**Computer Applications: Arts and Humanities**]: Linguistics

General Terms
Algorithms, Measurement, Performance, Human Factors

Keywords
Intelligent tutoring systems, natural language processing, corpus linguistics, data mining, automatic essay scoring, individual differences

1. INTRODUCTION
Automatically assessing writing quality is an important element of standardized tests such as the Graduate Record Exam (GRE) and the Test of English as Foreign Language (TOEFL), as well as intelligent tutoring systems such as the Writing-Pal (W-Pal [12]). Traditionally, automatic essay scoring (AES) systems have focused on textual features to assess writing quality. These features generally include linguistic elements related to word frequency, syntactic complexity, and cohesion along with discourse features related to text structure, theses, and topic sentences [2, 20]. AES systems have been quite successful, demonstrating strong correlations with human scores of essay quality. However, exact matches between raters and automatic scores have remained relatively low [2, 10, 15, 20]. One potential for increasing the accuracy of these systems is to consider the role of student attributes (e.g., demographic information, standardized text scores, and survey results) in statistical models of essay quality. Such individual differences may explain some of the variance that has not been captured by textual features.

Thus, this study examines the hypothesis that using student attributes in conjunction with linguistic and rhetorical elements of the text will increase the accuracy of automatic essay scoring. This study, by consequence, combines two contrasting research lines relating to writing quality. On the one hand, some writing researchers have focused on the relation between individual differences and essay quality [1, 15]. For example, more skilled readers [19], writers with stronger vocabularies [18], and writers with more writing-specific knowledge [16] are more likely to compose higher quality essays. On the other hand, most AES developers have tended to focus on the textual features that relate to essay quality, rather than on the prior skills or abilities of the writers, potentially because individual difference measures are either not available or are rarely collected in the context of AES.

Common tools used in the past to examine essay quality include the Biber tagger [3], Coh-Metrix [11], and the Writing Assessment Tool (WAT) [11]. The Biber Tagger automatically calculates features for lexical sophistication (e.g., type/token ratio and word length), cohesion and rhetorical features (e.g. conjuncts, hedges, amplifiers, and emphatics), grammatical features (e.g. nouns, verbs, nominalizations, and modals), and clause-level features (e.g. subordinations, complementation, and passives). Coh-Metrix calculates a number of text-based linguistic features related to lexical sophistication (word frequency, word concreteness, word familiarity, polysemy, hypernymy), syntactic complexity (incidence of infinitives, phrase length, number of words before the main verb), and cohesion (word overlap,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA
Copyright 2015 ACM 978-1-4503-3417-4/15/03…$15.00
http://dx.doi.org/10.1145/2723576.2723595

semantic similarity, incidence of connectives). WAT, in addition to reporting the above-mentioned Coh-Metrix and Biber Tagger features, measures n-gram frequency, rhetorical features, as well as additional measures of global cohesion and lexical and syntactic complexity. Text features such as these can be used to develop AES systems that automatically score students' essays based on their occurrence within the text. In general, such systems show correlations with human judgments of essay quality that range between .60 and .85. The scores from AES systems also report perfect agreement (i.e., exact matches between a human score and a score provided by the scoring system) from 30-60% and adjacent agreement (i.e., scores reported by the scoring system that are 1 point above or below the score provided by the human rater) from 85-99% [2, 10, 15, 20].

While AES systems show strong correlations and accuracy, some critics argue that the systems are impersonal, lack human sensitivity, and cannot respond to elements of writing quality that fall outside of the available algorithms [9]. Theoretically,, one would assume that including individual differences such as reading skill, vocabulary knowledge, and domain knowledge in AES models will not only make AES models more personal (and hence address some of the critics' concerns), but will also contribute to predictions of writing quality. Such a hypothesis is based on the notion that a student brings a constellation of skills to the composition process [8] and because individual attributes and essay features are assumed to serve as proxies both for the quality of the essay [10] and for individual differences [4]. We presume that including information about the writer's abilities as well as the features of the essay will better predict and provide more accurate estimates of the quality of an essay.

We examine this hypothesis in the context of the Writing Pal tutoring system [12]. W-Pal is an intelligent tutoring system designed to provide writing strategy instruction to high school and entering college students. Unlike AES systems, which focus on essay practice and sometimes provide support and instruction in the form of feedback (traditionally referred to as automatic writing evaluation, AWE, systems), W-Pal emphasizes strategy instruction first, followed by targeted strategy practice and then whole-essay practice. W-Pal provides instruction on writing strategies that cover three phases of the writing process: prewriting, drafting, and revising. Each of the writing phases is further subdivided into instructional modules. These modules include: *Freewriting* and *Planning* (prewriting); *Introduction Building*, *Body Building*, and *Conclusion Building* (drafting); and *Paraphrasing*, *Cohesion Building*, and *Revising* (revising). An important component of W-Pal is the incorporation of game-based practice. These games target specific strategies involved in the writing processes presented above (e.g., *freewriting, cohesion building, paraphrasing*) and provide students with opportunities to practice the strategies in isolation before moving on to practicing them during the composition of an entire essay.

Our overarching objective in this study is to enhance the accuracy of feedback within W-Pal. We accomplish this by using natural language processing tools to calculate linguistic and rhetorical features of essays which were composed by students who interacted with W-Pal. Additionally, we use the student attributes that we collected from these W-Pal participants. These linguistic features and student attributes were then included within a regression analysis to examine the extent to which they could predict essay quality (as scored by trained, expert raters). The purpose of this approach is to investigate the potential to use a combination of text features and student attributes to develop automatic scoring models and test their scoring accuracy. Such an approach has not been conducted in the past and may afford additional research avenues with which to automatically assess essay quality, provide summative feedback to users of AES systems, and provide new data mining approaches that can be used in tutoring systems, standardized writing assessments, and massive open on-line courses (MOOCs) to more accurately model writing proficiency.

2. METHODS

2.1 Participants

For this study, we recruited 87 students from public high schools in the metro Phoenix area. Students' average age was 15.6 years, with an average grade level of 10.4. Of the 87 participants, 62.1% were female and 37.9% were male. Thirty eight of the participants self-identified as English Language Learners (ELL). The remaining participants self-identified as native speakers of English (NS). Participants were divided into two conditions: the W-Pal condition (n = 42) or the Essay condition (n = 45). Of the 87 participants in both conditions, complete data for 86 of the participants was available. Additionally, posttest data from one student was not recorded due to a technical error. Therefore, we collected 171 pretest and posttest essays in total.

2.1 Procedures

Students attended 10 sessions (1 session/day) over a 2-4 week period. Participants wrote a pretest essay during the first session and a posttest essay during the last session. The essays were written on two prompts (on the value of competition and on the role of image) counterbalanced across the pretest and posttest essays. In addition, the first and final sessions included assessments of reading comprehension, vocabulary, writing proficiency, strategy knowledge, and writing attitudes (discussed below). Sessions 2-9 were devoted to training with students either interacting with W-Pal or the W-Pal automatic writing evaluation system. For this study, we used only the pretest and the posttest essays written by the students (N = 171). For the student attributes, we used only the data collected during the first session.

2.2 Essay Scoring

Each essay in the corpus was scored independently by two expert raters using a 6-point rating scale developed for the SAT (a college entrance exam common in the United States). The rating scale was used to holistically assess the quality of the essays and had a minimum score of 1 and a maximum score of 6. Raters were first trained to use the rubric with a small sample of argumentative essays. A Pearson correlation analysis was used to assess inter-rater reliability between raters. When the raters reached a correlation of $r = .70$, the ratings were considered reliable and the raters scored a larger subsection of the corpus. The final inter-rater reliability across all raters for all the essays in both corpora was $r > .70$. Average scores between the raters were calculated for each essay.

2.3 Student attributes

2.3.1 Demographic Information

Students' demographic information was collected at pretest. The demographic survey asked students to report basic information, such as their age, gender, average, grade point average, first language status, as well as their perceptions towards reading and writing. Additionally, the demographic survey assessed students'

performance orientation, as well as their comfort and excitement towards using computers.

2.3.2 Reading Comprehension

Students' reading comprehension ability was assessed through the Gates-MacGinitie Reading Skill test (4th Ed.; Form S; level 10/12 [7]). The test consists of 48 multiple-choice questions that measure students' ability to comprehend both shallow and deep level information across 11 short passages. In the current study students' comprehension scores on this test ranged from 10 to 45 (M=24.59, SD=8.91).

2.3.3 Vocabulary Knowledge

Students' vocabulary knowledge was measured through the use of the Gates-MacGinitie Reading *S*kill test (4th *E*d.*; F*orm S*;* level 10/12 [7]*)*. This test assesses vocabulary skill by showing students 45 sentences or phrases that each contain an underlined vocabulary word and ask the students to select a word from a list of 5 that is most closely related to the underlined word. In the vocabulary portion of this experiment, students' scores ranged from 6 to 45 (M=26.63, SD=8.89).

2.3.4 Writing Apprehension

Students' apprehension toward writing was measured with the Daly-Miller Writing Apprehension Test (WAT) [6]. The Daly-Miller WAT assesses an individual's level of apprehension toward writing. This assessment includes items related to evaluation apprehension (fear of evaluation), stress apprehension (general fear of writing manifesting early in the writing process), and product apprehension (fear of writing manifesting as a general disdain for writing).

2.3.5 Prior Knowledge

Prior knowledge was measured with a 30-question assessment that was designed for high school students. It has been previously used in research related to strategy training and reading comprehension [13, 14]. This measure assesses knowledge in the domains of science, literature, and history.

2.3.6 WASSI

The Writing Attitudes and Strategies Self-Report Inventory (WASSI) is a self-report measure that was administered to test students' writing attitudes and strategy use. Students respond to statements about themselves by indicating their level of agreement on a 6-point scale ranging from strongly disagree to strongly agree. The WASSI comprises four different subscales (prewriting, drafting, attitudes, and self-efficacy) each targeting a different aspect important to writing performance.

2.4 Text Features

Linguistic features from the text were computed using Coh-Metrix and the Writing Assessment Tool (WAT). These features are discussed briefly below. More detailed descriptions for the tools and the features on which they report can be found in [5, 10, 11].

2.4.1 Coh-Metrix

Coh-Metrix represents the state of the art in computational tools and is able to measure text difficulty, text structure, and cohesion through the integration of lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics. Coh-Metrix reports on linguistic variables that are primarily related to text difficulty. These variables include indices of causality, cohesion (semantic and lexical overlap, lexical diversity, along with incidence of connectives), part of speech and phrase tags (e.g., nouns, verbs, adjectives), basic text measures (e.g., text, sentence, paragraph length), lexical sophistication (e.g., word frequency, familiarity, imageability, familiarity, hypernmy, concreteness), and syntactic complexity (e.g., words before the main verb, noun phrase length, and incidence of infinitives). These Coh-Metrix indices have been used successfully in a number of studies that focus on predicting essay quality [5, 10, 11]. For additional information about the types of indices calculated by Coh-Metrix and how the calculations are made, we refer the reader to [11].

2.4.2 WAT

WAT computes linguistic features specifically developed to assess student writing. These features include indices related to global cohesion, topic development, n-gram accuracy, lexical sophistication, key word use, and rhetorical features. Cohesion features include LSA measures between paragraph types (introduction, body, and conclusion paragraphs) and LSA measures of relevance. N-gram accuracy features include indices related to n-gram frequency, n-gram proportion, and correlations between expected and actual n-gram use (at the level of bi-grams and tri-grams calculated on written and spoken corpora). Rhetorical features include indices such as hedges, conjuncts, amplifiers, and conclusion statements. The features reported by WAT have been used in a number of studies that have successfully investigated links between essay quality and textual features [5, 10]. For additional information about the types of indices calculated by WAT and how the calculations are made, we refer the reader to [10].

2.5 Analyses

Prior to the regression analysis, correlation analyses were conducted to examine the strength of relations between the selected indices and the human scores of essay quality. If an index demonstrated a significant correlation and at least a small effect size with the human scores (r > .100), it was retained in the analysis. Multicollinearity was then assessed between the indices (r > .900). When two or more indices demonstrated multicollinearity, we retained the index that correlated more strongly with the scores of essay quality. Lastly, variables that were not normally distributed were removed from the analyses. A training and test set were used in the regression analysis to ensure that the results were generalizable to the population. The training set was comprised of approximately 67% of the essays while the test set was comprised of approximately 33% of the essays. Exact and adjacent accuracy is also reported for the scores calculated by the resulting regression model. Exact matches demonstrate perfect agreement between human and regression scores while adjacent agreement occur when human and automated scores are within one point of each other.

3. Results

3.1 Correlations and Normality Checks

Of the 292 selected features, 106 of the features demonstrated significant correlations and at least a small effect size with essay scores. Of these 106 variables, 11 of the variables demonstrated multicollinearity with other variables that correlated more strongly with the essay quality scores. These 11 variables were removed. Of the remaining 95 variables, 25 of the variables were not normally distributed and were removed. This trimming of variables left us with 70 features with which to predict essay quality. Of these 70 variables, 60 were textual features and 10

Table 1: Linear regression results for student attributes and textual features

Entry	Variable Added/Removed	Correlation	R-Squared	*B*	SE	B
Entry 1	Frequency spoken bigrams	0.596	0.355	-77.223	9.49	-0.488
Entry 2	Word concreteness	0.689	0.474	-0.004	0.001	-0.193
Entry 3	GMRT Vocabulary	0.735	0.54	0.02	0.004	0.282
Entry 4	LSA body to conclusion paragraphs	0.757	0.573	21.528	6.263	0.205
Entry 5	Noun hypernymy	0.771	0.595	0.222	0.078	0.166
Entry 6	Incidence of infinitives	0.783	0.614	0.079	0.033	0.146

Notes: Estimated Constant Term is 3.165; *B* is unstandardized Beta; SE is standard error; B is standardized Beta

were student attributes including 1 demographic attribute, 2 standardized test scores, and 7 variables from survey answers.

3.2 Multiple Regression:

3.2.1 Regression Analysis Training Set

The linear regression using the selected variables yielded a significant model, $F(6, 117) = 30.977$, $p < .001$, $r = .783$, $r^2 = .614$. Six variables were significant predictors in the regression: frequency of spoken bi-grams, word concreteness, Gates-MacGinitie (GMRT) vocabulary test scores, LSA similarity between body and conclusion paragraphs, noun hypernymy, and the incidence of infinitives. Five of the variables were text-based while one variable (GMRT vocabulary) was related to student attributes. The remaining variables were not significant predictors and were not included in the model. The regression model is presented in Table 1. The results from the linear regression demonstrate that the combination of the six variables accounts for 61% of the variance in the human judgments of writing quality.

3.2.2 Regression Analysis Test Set

The model for the test set yielded $r = .698$, $r^2 = .487$. The results from the test set model demonstrate that the combination of the six variables accounted for 49% of the variance in the evaluation of the 49 essays comprising the test set.

3.2.3 Exact and Adjacent Matches

The regression model produced exact matches between the predicted essay scores and the human scores for 115 of the 171 essays (67% exact accuracy). The model produced adjacent matches for 169 of the 171 essays (99% adjacent accuracy).

4. Discussion

We have investigated the potential for student attributes in combination with textual features to predict writing quality. The findings demonstrate that a combination of text features and student attributes significantly predicts human ratings of essay quality. Such findings indicate that student attributes can be used to increase the accuracy of scoring models and that their inclusion in scoring models may open up new avenues for improving the personalization of the scoring models. Specifically, the use of personalized data may allow tutoring and AES/AWE systems to provide more effective feedback to students beyond text features alone. By incorporating student information into essay scoring models, the formative feedback in these writing systems can focus on the needs of the individual students rather than on individual essays. Additionally, the use of this student data may allow automated writing systems to adapt lessons to users based on their writing performance and individual characteristics.

The regression model included indices related to both student and text features, providing evidence that a combination of both features leads to gains in scoring accuracy (as hypothesized). The model reported was slightly higher than the expected adjacent accuracy range as found in previous published models (i.e., 65%) and reported an exact accuracy that was also higher than previous models (i.e., 99%). However, the reported correlations to human scores were on the lower end of acceptability when compared to previous models. Textually, the model demonstrated that higher quality essays were marked by the use of more infrequent bigrams, less concrete words, greater semantic similarity between body and conclusion paragraphs, fewer specific words, and greater use of infinitive clauses. From a student attribute perspective, higher quality essays were written by students with greater vocabulary knowledge, supporting previous research [1, 18]. The results from the regression model demonstrate that both student and text features can act in unison to provide accurate essay scores for students who use the W-Pal system. The inclusion of both student and textual features should help improve feedback mechanisms and increase the validity of the W-Pal AES system.

The model reported that the strongest predictors of essay quality were text features. This may indicate that text features are more important elements of essay quality; however, these results may also be reflective of our data collection methods, which limited the number of student attributes we could include in the statistical modeling. There were a number of student attributes that were not included in the current data collection that could be collected in future studies. These could include: simple survey questions related to writing strategy use, socio-economic status, future educational plans, amount of writing or reading completed at home or in the students' free time, and specific grades in specific classes (to name but a few). Standardized test scores related to writing ability, math skills, and content knowledge could also be included in the models, as could working memory ability. In addition, the current models do not take into consideration sequential information such as the students' score on their previous essay or the students' scores on the games and quizzes included in each W-Pal writing module. Such scores could be used to provide an updated model of the students' current knowledge and help improve overall scoring accuracy.

In addition, the models could be improved with the inclusion of additional text features and different statistical analyses. From a text-based perspective, the models discussed above did not include indices related to specific discourse units. Such discourse units would include the presence and the strength of items such as

thesis statements, arguments, topic sentences, and supporting evidence. Indices that measure these elements are in the process of development and will be added to WAT in the near future Similarly, the current version of WAT does not calculate indices related to grammatical and mechanical (i.e., spelling and punctuation) accuracy. These indices (also under development) may add to the ability for the W-Pal AES system to more accurately assign scores to essays. Statistically, the model reported in this analysis takes a linear approach to essay scoring. Other approaches including a hierarchical approach or a clustering approach might also lead to improved results [5].

5. Conclusion

Overall, the findings from this study provide evidence for the use of both text features and student attributes in conjunction to improve automatic essay scoring. The results also provide an indication of which features are most predictive of writing quality, providing a snapshot of how text features affect human judgments of writing quality and how student attributes relate to writing success. These findings have important implications for both educators and researchers because they reveal that essay scoring approaches should incorporate measures not only about the essay itself, but also about the writer. The inclusion of such indices may improve the accuracy of the scores assigned to essays, as well as increase the validity and the personal nature of the feedback provided to students on their writing. Future studies should focus on a greater number of text features and student attributes, writing samples written outside of a tutoring system, and writing samples that are based on other genres such as integrated writing or content based writing.

6. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We also thank Rod Roscoe, Tanner Jackson, and Jianmin Dai for their help.

7. REFERENCES

[1] Allen, L. K., Snow, E. L., and 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.

[2] Attali, Y. and Burstein, J. 2006. Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*. *4*, 3

[3] Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

[5] Crossley, S. A., Roscoe, R., and McNamara, D. S. 2014. What is quality writing? An investigation into the multiple ways writers can write high quality essays. *Written Communication, 31* (2), 184-214.

[6] Daly, J.A., and Miller, M.D. 1975. Apprehension of writing as a predictor of message intensity. *Journal of Psychology, 89*, 175-177.

[7] *Gates-MacGinitie Reading Tests.* 1989. *Technical Report for Gates-MacGinitie Reading Tests form S.* Chicago, Illinois*:* Riverside *Publishing.*

[8] Graham, S. 2006. Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457-477). Mahwah, NJ: Erlbaum.

[9] Hearst, M. 2002. The debate on automated essay scoring. *Intelligent Systems and their Applications, IEEE*, *15*, 22-37.

[10] McNamara, D., Crossley, S., and Roscoe, R. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavioral Research Methods, Instruments and Computers, 45,* 499-515.

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

[12] McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., and Graesser, A. 2012. The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298-311). Hershey, P.A.: IGI Global.

[13] O'Reilly, T. and McNamara, D.S. 2007. The impact of science knowledge, reading strategy knowledge on more traditional "High-Stakes" measures of high school students' science achievement. *American Educational Research Journal, 44,* 161-196.

[14] O'Reilly, T., Best, R., and McNamara, D.S. 2004. Self-explanation reading training: Effects for low-knowledge readers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1053-1058). Mahwah, NJ: Erlbaum.

[15] Rudner, L., Garcia, V., and Welch, C. 2006. An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, *4*, 4.

[16] Saddler, B., and Graham, S. 2007. The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading and Writing Quarterly, 23,* 231-247.

[17] Scardamalia, M., and Bereiter, C. 1987. Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Reading, writing, and language learning* (vol. 2, pp. 142-175). New York: Cambridge University Press.

[18] Stæhr, L. S. 2008. Vocabulary size and the skills of listening, reading and writing. Language Learning Journal, 36(2), 139-152.

[19] Tierney, R. J., and Shanahan, T. 1991. Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, ppp. 246-280). New York: Longman

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