Toward Revision-Sensitive Feedback in Automated Writing Evaluation

Rod D. Roscoe
Arizona State University
Rod.Roscoe@asu.edu

Matthew E. Jacovina
Arizona State University
Matthew.Jacovina@asu.edu

Laura K. Allen
Arizona State University
LauraKAllen@asu.edu

Adam C. Johnson
Arizona State University
acjohn17@asu.edu

Danielle S. McNamara
Arizona State University
Danielle.McNamara@asu.edu

ABSTRACT

Revising is an essential writing process yet automated writing evaluation systems tend to give feedback on discrete essay drafts rather than changes across drafts. We explore the feasibility of automated revision detection and its potential to guide feedback. Relationships between revising behaviors and linguistic features of students' essays are discussed.

Keywords

Automated Writing Evaluation; Writing; Revising; Intelligent Tutoring Systems; Natural Language Processing; Feedback

1. INTRODUCTION

Automated writing evaluation (AWE) systems provide computer-based scores and feedback on students' writing, and can promote modest gains in writing quality [1, 2]. One concern is that students receive feedback on their *current* drafts that ignores *patterns of change* from draft to draft. We argue AWE tools should include feedback models that incorporate data on students' revising behaviors and textual changes. These innovations may afford greater personalization of formative feedback that helps students recognize how their editing actions affect writing quality.

This study used Writing Pal (W-Pal), a tutoring and AWE system that supports writing instruction and practice [3, 4]. When submitting essays to W-Pal, students receive scores (6-point scale) and feedback with actionable suggestions for improvement. Scoring and feedback are driven by natural language processing (NLP) algorithms that evaluate lexical, syntactic, semantic, and rhetorical text features [1, 5]. One goal for W-Pal development is feedback that promotes more effective revising [see 4].

2. METHOD

2.1 Context and Corpus

High school students (n = 85) used W-Pal to write persuasive essays on the topic of "fame." Most identified as native English speakers (56%) and others as English-language learners (44%).

2.2 Detection and Annotation of Revising

We calculated difference scores between drafts for several NLP measures (via Coh-Metrix [5, 6]). Lexical measures assessed word choice and vocabulary, such as word frequency and hypernymy. Cohesion indices assessed factors such as overall essay cohesion, semantic relatedness (using LSA), and structure.

Human annotation of revisions adapted methods from prior research [7, 8]. Writers can alter their text via adding, deleting, substituting, or reorganizing actions. Human coding of these revision actions showed high reliability (κ = .92). Revisions can also maintain (superficial edits) or transform (substantive edits) the meaning of surrounding text. Human coding of revision impact on text meaning also demonstrated high reliability (κ = .81).

3. RESULTS

3.1 Automated Detection of Revising

Essays demonstrated detectable changes in linguistic features from original to revised drafts. Revised essays were longer, included more transitional phrases and first-person pronouns, and were somewhat more cohesive (see Table 1).

Table 1. Linguistic Changes and Correlations with Scores

	Linguistic Change		Correlation with Score Change	
Linguistic Change	***t*(84)**	***p***	***r*(84)**	***p***
Basic				
Word Count	**6.24**	**< .001**	.06	.593
Sentence Count	**4.33**	**< .001**	-.09	.393
Lexical				
Lexical Diversity	-0.28	.781	.17	.124
Word Concreteness	0.83	.410	**.34**	**.002**
Word Familiarity	-0.74	.463	-.01	.954
Word Hypernymy	0.80	.424	**.24**	**.028**
1st Person	**2.09**	**.040**	-.07	.545
2nd Person	-1.06	.294	**-.22**	**.043**
3rd Person	-0.23	.818	-.10	.342
Cohesion				
Connectives	**1.67**	**.099**	.03	.809
LSA Given/New	**2.98**	**.004**	.08	.484
LSA Sentences	0.58	.562	**.24**	**.029**
LSA Paragraphs	**1.86**	**.066**	-.08	.465
Deep Cohesion	0.71	.478	**.18**	**.098**
Referential Cohesion	0.52	.607	.01	.893
Narrativity	1.05	.296	**-.25**	**.023**

Essay quality increased from original (M = 2.7, SD = 1.0) to revised drafts (M = 2.9, SD = 1.1), $t(84) = 3.64$, $p < .001$, $d = .19$. Gains correlated with increased concreteness, specificity, objectivity (i.e., fewer 2nd-person pronouns and less story-like), and cohesion. Importantly, the linguistic changes linked to gains were *not* the most typical changes. This finding reinforces the idea that students are not skilled revisers—their revising behaviors can be dissociated from actions that improve the quality of their work.

3.2 Human Annotation of Revising

The most common revisions were additions (47.5%) and substitutions (33.6%). Deletions (15.4%) and reorganizations (2.5%) occurred less often. None of the revising actions were correlated with changes in essay score. This finding reiterates the point that high school students are not necessarily skilled revisers.

3.3 Relationships between Modes of Analysis

The total number of revisions was not related to linguistic changes across drafts (range of *r*s from -.18 to .12). Simply revising *more* had minimal effects. Additions, substitutions, and reorganization had few effects. In contrast, deletions were associated with reductions in narrativity and third-person pronouns. Along with reduced word familiarity, this pattern suggests that students were removing story-like language. Deletions were also associated with reduced given information, semantic similarity across paragraphs, and referential cohesion. Thus, as students removed content from their essays, the cohesive flow of ideas was perhaps hindered. Overall, deletions seemed to be linked to both gains and setbacks in essay quality (see Table 2).

Table 2. Correlations of Revision Types and Linguistic Change

Linguistic Change	Add	Delete	Subst.	Reorg.
Basic				
Word Count	**.29**[b]	**-.36**[a]	-.18	-.10
Sentence Count	**.37**[a]	-.18	-.16	.05
Lexical				
Lexical Diversity	.01	**.26**[c]	-.04	.07
Word Concreteness	.00	**.29**[b]	.08	.06
Word Familiarity	-.04	**-.28**[c]	.15	-.09
Word Hypernmy	-.10	.11	.02	-.18
1st Person	.04	-.11	.11	.07
2nd Person	-.09	-.03	-.05	-.04
3rd Person	-.01	**-.26**[c]	-.07	.00
Cohesion				
Connectives	-.07	.16	.09	-.03
LSA Given/New	-.02	**-.32**[c]	-.07	-.07
LSA Sentences	-.20	-.09	.06	-.12
LSA Paragraphs	.07	**-.24**[c]	-.05	.04
Deep Cohesion	.00	-.11	.07	-.07
Referential Cohesion	-.10	**-.25**[c]	.12	-.03
Narrativity	-.07	**-.34**[a]	-.01	.01

Note. [a]$p \leq .001$. [b]$p \leq .01$. [c]$p \leq .05$.

A final analysis examined revisions by both type *and* impact. As in the previous analysis, the most meaningful linguistic changes were associated with deletions, with substantive deletions appearing to have the strongest influence. Superficial deletions tended to make essays more personalized (i.e., more 1st-person pronouns) and less specific. Substantive deletions tended to make essays shorter, less story-like, more sophisticated in terms of vocabulary, and less cohesive.

4. Discussion

Our results provide evidence that automated tools can detect linguistic changes in students' writing. Formative feedback based on such measures might help students appreciate when and how their drafts evolve over time. For instance, when an increase in narrativity or decrease in cohesion are detected, feedback could flag the edited sections of text so that conscientious students can draw inferences about the impact of their revisions.

Ideally, AWEs should also be able to detect and give feedback on revising behaviors. From the current study, however, it is unclear whether linguistic data could be used to identify such behaviors. With the exception of deletions, students' revising actions did not have a profound impact on linguistic properties.

One solution may reside in keystroke logging [9]. Keyboard and mouse clicks made while interacting with an AWE system may be interpretable with respect to revising. For example, backspace presses may indicate deletion. The use of mouse buttons to select text, along with "CTRL-X" and "CTRL-V" hotkey functions, could signal reorganization. If such tools can be added to AWEs, they may provide real-time measures of writing and revising behaviors that can be explicitly linked to linguistic consequences.

5. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Educational Sciences (IES R305A120707). Opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the IES.

6. REFERENCES

[1] Shermis, M., and Burstein, J. C. (Eds). 2013. *Handbook of automated essay evaluation: current applications and new directions.* Routledge.

[2] Stevenson, M., and Phakiti, A. 2013. The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65.

[3] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., and McNamara, D. S. 2014. The Writing Pal intelligent tutoring system: usability testing and development. *Computers and Composition*, *34*, 39-59.

[4] Roscoe, R. D., Snow, E. L., Allen, L. K., and McNamara, D. S. 2015. Automated detection of essay revising patterns: applications for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning*, *10*, 59-79.

[5] 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.

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

[7] Faigley, L., and Witte, S. 1981. Analyzing revision. *College Composition and Communication*, *32*, 400-414.

[8] Fitzgerald. J. 1987. Research on revision in writing. *Review of Educational Research*, *57*, 481-506.

[9] Leijten, M., and Van Waes. 2013. Keystroke logging in writing research: using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358-392.