Predicting Misalignment between Teachers' and Students' Essay Scores Using Natural Language Processing Tools

Laura K. Allen[1], Scott A. Crossley[2], and Danielle S. McNamara[1]

[1] Learning Sciences Institute, Arizona State University, Tempe, AZ 85287
laurakallen@asu.edu, dsmcnamra1@gmail.com

[2] Department of Applied Linguistics/ESL, Georgia State University, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, USA
scrossley@gsu.edu

Abstract. We investigated linguistic factors that relate to misalignment between students' and teachers' ratings of essay quality. Students (n = 126) wrote essays and rated the quality of their work. Teachers then provided their own ratings of the essays. Results revealed that students who were less accurate in their self-assessments produced essays that were more causal, contained less meaningful words, and had less argument overlap between sentences.

Keywords: Cohesion, Intelligent Tutoring Systems, Natural Language Processing, Corpus Linguistics, Computational Linguistics, Writing Pedagogy

1 Introduction

One factor that is important for students' writing proficiency is their ability to monitor their own performance [1]. Importantly, the accuracy of this monitoring is a key component to successfully navigating any learning task. When students are aware of how well they are performing, they can more carefully select their learning goals and behaviors, which consequently leads to better performance and retention [2].

Previous studies have reported that students' ratings of their own essay performance are largely divergent from the ratings provided by their teachers or other expert raters [1; 3]. This indicates that there may be a "breakdown" in the link between students' understanding of their own performance and more objective criteria for quality writing. Varner and colleagues (2013) referred to these differences as *evaluative misalignment* [3]. They suggested that students may struggle to produce high-quality texts because their criteria for quality rating are not in line with those of their teachers. As a result, they may produce essays that do not meet the standards set by their teachers and not understand why they receive the scores that they do.

The current study investigated the linguistic factors that relate to the degree of misalignment between students' and teachers' ratings of essay quality. We first examine whether there are specific linguistic features of students' essays that predict the mag-

nitude of their misalignment from the teachers. We then conduct correlations to determine whether and how these indices relate to students' and teachers' essay ratings.

2 Methodology

The participants were high school students (n=126) enrolled in tenth-grade English courses. They wrote 25-minute essays as practice for the writing portion of the SAT and were asked to rate the quality of their essays on a scale from 1-6. Additionally, teachers rated the essays from 1-6. Linguistic features of the essays were calculated to identify misalignment between the teachers' and the students' essay scores and to assess relations between the essay scores and these linguistic features.

2.1 Student and Teacher Essay Ratings

Students assigned their essays an average score of 4.04 (*SD*=0.81), whereas teachers displayed an average rating of 3.67 (*SD*=1.01). Thus, students tended to overestimate their essay ratings; $t(125)=3.86$, $p<.001$, Cohen's $d=.40$. Further, the student and teacher ratings were only moderately correlated ($r=.26$, $p<.01$). These results suggest a possible misalignment between students' and teachers' criteria for writing quality.

2.2 Selected Linguistic Features

To examine the linguistic features that were predictive of misalignment, we used three natural language processing tools: Coh-Metrix [4], TAALES [5], and TAACO [6]. We selected linguistic features that fell into four categories: Text length indices, syntactic complexity indices, lexical sophistication indices, and cohesion indices. We refer the reader to 4, 5, and 6 for additional information. We removed all selected indices from the analysis that lacked normal distributions.

2.3 Statistical Analysis

For each student, a misalignment score was calculated by using the absolute value of the difference between the student's self-assessment and the teacher's essay rating. The scores were placed into three categories: *aligned* (i.e., the student and teacher assigned the same grade), *misaligned by 1* (i.e., difference of 1 between scores), or *misaligned by 2 or greater* (i.e., differences in score 2 or greater).

3 Results

3.1 Group prediction

A MANOVA was conducted using the linguistic indices as the dependent variables and the misalignment groups as the independent variables. Sixteen variables related to

lexical sophistication and cohesion demonstrated significant differences between the groups while not demonstrating multi-collinearity with each other.

A stepwise discriminant function analysis (DFA) retained three of these variables related to lexical sophistication and cohesion as significant predictors of whether essay scores were aligned, misaligned by 1, or misaligned by 2 or greater. These variables were *causal verbs*, *word meaningfulness*, and *part of speech overlap between adjacent sentences*. Results demonstrate that the DFA using these three indices correctly allocated 69 of the 126 texts in the total set, $\chi 2$ (df=1, n=126) = 25.022 p < .001, for an accuracy of 54.8%. The reported Cohen's Kappa was .310, indicating a fair agreement. For the leave-one-out cross-validation (LOOCV), the discriminant analysis also allocated 66 of the 126 texts for an accuracy of 52.4%. The accuracy of these DFA analyses did not vary for low- and high-quality essays. For essays that were rated between 1-3 by the teachers, the model accuracy was 54.7% and for essays that were rated between 4-6, the model accuracy was 54.8%.

3.2 Correlations with student and teacher scores

Correlations were conducted between the selected indices from the DFA and the essay scores assigned by the students and by the teachers. Neither students' nor teachers' scores correlated strongly with indices related to text length, syntactic complexity, or lexical sophistication. However, a number of the cohesion indices demonstrated small effects. For the student scores, four indices demonstrated at least a small (but not significant) effect: *adjacent overlap three sentences content words*, *adjacent overlap one sentence POS tags*, *adjacent overlap two sentences nouns*, and *adjacent overlap two sentences all words*. This analysis revealed that student scores were negatively related to all three of these indices, except for *adjacent overlap two sentences nouns*. For the teacher scores, four indices demonstrated at least a small effect size (only one index was significant): *adjacent overlap one sentence lemma*, *causal verbs*, *adjacent overlap two sentences all words*, and *LSA paragraph to paragraph standard deviation*. This analysis revealed that the teachers' scores were also related to both local and global cohesion indices, albeit different indices than the students' scores.

4 Discussion

The results of our DFA analyses revealed that student and teacher misalignments were, indeed, systematically related to specific linguistic features in the essays written by the students. Students with misalignments produced essays that were more causal, contained less meaningful words, and had less argument overlap between sentences. In other words, these students produced more narrative texts (i.e., causal texts) that contained more difficult words, and less local cohesion. These results may suggest that the students and teachers in this study varied in their sensitivity to certain linguistic properties, which may have driven them to assign different ratings to the essays.

The correlational results indicated that differences in the students' and teachers' essay scores were most apparent at the cohesion levels. These results potentially suggest that the teachers were more aware of the nuances related to essay cohesion, whereas

the students may have simply perceived all cohesion indices to be similarly (i.e., negatively) associated with quality. More importantly, however, the results of the correlation analyses revealed that the linguistic indices that were predictive of student-teacher misalignment were different than the linguistic indices that predict essay quality (from both the student and teacher perspective). Previous research studies have shown that linguistic features of students' essays are related to student and teacher ratings of essay quality [7]. However, in the current study, these variables were not the same variables that were predictive of misalignment. This suggests that the properties of essays that may contribute to perceptions of essay quality are different than those that lead students to make inaccurate assessments of their own performance.

Overall, the results from this study suggest that students' difficulties with monitoring performance may stem, at least in part, from their misunderstandings of the criteria for quality writing. Additionally, they suggest that natural language processing tools can provide more fine-grained information related to these differences.

5 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.

6 References

1. Varner L.K., Roscoe, R.D., McNamara, D.S.: Evaluative Misalignment of 10th-Grade Student and Teacher Criteria for Essay Quality: An Automated Textual Analysis. Journal of Writing Research. 5, 35-59 (2013)
2. Dunlosky, J., Ariel, R.: Self-Regulated Learning and the Allocation of Study Time. In B. Ross (Ed.), Psychology of Learning and Motivation, pp. 103-140 (2011)
3. Kos, R., Maslowski, C. Second Graders' Perceptions of What is Important in Writing. The Elementary School Journal. 101, 567-585 (2001)
4. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, Cambridge (2014)
5. Kyle, K., Crossley, S.A.: Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly (in press)
6. Crossley, S.A., Kyle, K., McNamara, D.S. Automatic Assessment of Local and Global Cohesion: Implications for Text Comprehension, Coherence, and Quality. Discourse Processes (under review)
7. McNamara, D.S., Crossley, S.A., Roscoe, R.D., Allen, L.K., Dai, J. Natural Language Processing in a Writing Strategy Tutoring System: Hierarchical Classification Approach to Automated Essay Scoring. Assessing Writing. 23, 35-59 (2015)