Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System

Scott A. Crossley[1], Laura K. Varner[2], Rod D. Roscoe[2], and Danielle S. McNamara[2]

[1] Department of Applied Linguistics/ESL, Georgia State University,
34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, USA
scrossley@gsu.edu
[2] Learning Sciences Institute, Arizona State University, Tempe, AZ 85287
{laura.varner,rod.roscoe}@asu.edu, dsmcnamra1@gmail.com

Abstract. We present an evaluation of the Writing Pal (W-Pal) intelligent tutoring system (ITS) and the W-Pal automated writing evaluation (AWE) system through the use of computational indices related to text cohesion. Sixty-four students participated in this study. Each student was assigned to either the W-Pal ITS condition or the W-Pal AWE condition. The W-Pal ITS includes strategy instruction, game-based practice, and essay-based practice with automated feedback. In the ITS condition, students received strategy training and wrote and revised one essay in each of the 8 training sessions. In the AWE condition, students only interacted with the essay writing and feedback tools. These students wrote and revised two essays in each of the 8 sessions. Indices of local and global cohesion reported by the computational tools Coh-Metrix and the Writing Assessment Tool (WAT) were used to investigate pretest and posttest writing gains. For both the ITS and the AWE systems, training led to the increased use of global cohesion features in essay writing. This study demonstrates that automated indices of text cohesion can be used to evaluate the effects of ITSs and AWE systems and further demonstrates how text cohesion develops as a result of instruction, writing, and automated feedback.

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

1 Introduction

For many students, developing writing proficiency is a challenging [1] yet crucial aspect of academic and professional success [2]. To facilitate such writing development, research has emphasized both the teaching of writing strategies [3] and providing students with formative feedback on how to improve writing [4]. For example, local and global cohesion are key linguistic properties of a text that may contribute to the readability and coherence of a text [5-6]. Knowing this, composition instructors might teach students strategies for building cohesion and might offer feedback about "awkward transitions" or "non sequiturs" (i.e., cohesion breaks) in students' written

K. Yacef et al. (Eds.): AIED 2013, LNAI 7926, pp. 269–278, 2013.

work. Such pedagogical principles for strategy instruction and feedback can also be implemented within computer-based technologies for writing instruction, such as intelligent tutoring systems (ITSs) and automated writing evaluation (AWE) systems. The Writing Pal (W-Pal) [7] tutoring system offers strategy instruction and game-based practice across multiple aspects of the writing process. W-Pal also allows students to author original prompt-based essays, which are scored and receive feedback guided by natural language processing (NLP) algorithms.

In W-Pal, and related computer-based systems for writing instruction, automated assessment is a fundamental ingredient of success. NLP algorithms are necessary to detect or diagnose particular strategies or writing errors, such as students' use or omission of cohesive cues. Likewise, algorithms inform the assessment of students' overall writing proficiency or growth. In this study, our goal is to investigate automated indices of cohesion as potential measures of writing growth. This investigation occurs within the context of W-Pal, and uses a variety of automated features of cohesion found in the computational tools Coh-Metrix [8] and the Writing Assessment Tool (WAT) [9]. We specifically examine indices of local cohesion (i.e., connections between smaller text elements, such as sentences) and global cohesion (i.e., connections between larger text elements, such as paragraphs). These indices are employed to contrast writing development across two groups of writers. One group interacted with the complete W-Pal ITS, including strategy instruction, game-based practice, and essay-based practice with automated feedback. A second group used only the essay-based practice and feedback components of W-Pal, but wrote twice as many essays. Our hypothesis is that interacting with the complete W-Pal ITS will lead to the increased use of cohesive devices in student writing over time.

1.1 Cohesion

Cohesion refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text. Cohesion is contrasted with *coherence*, which refers to the understanding that the reader derives from the text. This coherence may be dependent on a number of factors, including linguistic features, background knowledge, and reading skill [10]. Pedagogically, text cohesion is a common theme in writing research [5] and textbooks [6]. Pedagogical perspectives promote the idea that the use of cohesive features in essays increases writing quality. However, empirical support for such assumptions has been mixed.

In two studies, Crossley and McNamara [11-12] investigated the degree to which analytical rubric scores of essay quality (e.g., essay coherence, strength of thesis) predicted holistic essay scores. Results of both studies found that human judgments of text coherence were the most informative predictors of human judgments of essay quality. However, neither of the studies found strong correlations between computational indices of local cohesion (e.g., indices of causal cohesion, spatial cohesion, temporal cohesion, connectives, and word overlap) and human judgments of text coherence. Crossley and McNamara [12], however, found that automated indices of global cohesion (LSA vector between paragraphs) correlated strongly with human judgments of coherence in essays. These studies suggest that *local* cohesive devices may not underlie the development of coherent textual representations of essay quality, but that *global* cohesive devices may contribute.

As measures of writing proficiency rather than text coherence, there are some indications that cohesion features are important in predicting human judgments of essay quality. McNamara et al. [9] found that a cohesion feature related to given information was positively predictive of essay quality. For counterexamples, however, see [13-14], and [9], which demonstrated that cohesion features may not correlate with human ratings or may correlate negatively with such judgments.

1.2 Automated Writing Evaluation

AWE systems provide opportunities for students to practice writing and receive holistic scores and feedback (i.e., deliberate practice) in the absence of a teacher. Deliberate practice is an important aspect of writing development. Like trained musicians and athletes, writers gain from extended practice [15-16] because such practice promotes self-regulation of planning, text generation, and reviewing [16]. However, deliberate practice also requires timely and relevant feedback. In writing instruction, such feedback may be provided by AWE systems, which reduce burdens placed on instructors and offer writers more opportunities to practice writing [17]. The algorithms that underlie AWE systems generally provide accurate scores to users, reporting perfect agreement of 30-60% and adjacent agreement of 85-99% [9, 18].

AWE systems have been critiqued for a variety of reasons. For instance, the scoring reliability of many AWE systems has recently been criticized [18], as has the potential for AWE systems to overlook infrequent writing problems that, while rare for a majority of writers, may be frequent to an individual writer. Such errors will likely not be assessed in an AWE system. Lastly, AWE systems have been criticized for depending on summative feedback at the expense of formative feedback [19].

1.3 The Writing Pal

ITSs that focus on teaching writing strategies adopt a pedagogical focus and are an alternative to strict AWE systems, although they often include AWE systems. W-Pal [7] is an ITS that adopts such a pedagogical focus. Unlike an AWE system that would focus only on essay practice with some supportive instruction, W-Pal emphasizes strategy instruction and targeted strategy practice prior to whole-essay practice. This strategy instruction is intended to facilitate task performance and accelerate skill acquisition and the acquisition of learning strategies, all of which are effective at improving student writing, particularly for adolescent writers [3].

W-Pal teaches writing strategies that cover three phases of the writing process. Each of the writing phases is subdivided into instructional modules: *Freewriting* and *Planning* (prewriting phase); *Introduction Building*, *Body Building*, and *Conclusion Building* (drafting phase); and *Paraphrasing*, *Cohesion Building*, and *Revising* (revising phase). An important component of W-Pal is that it incorporates a suite of games that target specific strategies. The games allow students to practice the strategies in isolation before applying the strategies to the essay writing process. The essay writing component of the system allows students to compose essays and then provides holistic scores and automated, formative feedback based on natural language input.

This feedback depends on the W-Pal AWE system, which focuses on strategies taught in the W-Pal lessons (including cohesion strategies). Thus, within W-Pal, students first view lessons that teach individual strategies; they then practice these strategies via games; lastly, they write practice essays for each of the modules and receive automated feedback from the AWE system on the quality of these essays.

2 Methodology

We collected writing data from two groups of students. The first group interacted with the full W-Pal system described above. The second group wrote and revised essays based only on feedback from the W-Pal AWE system. Both groups wrote pretest and posttest essays. We selected the W-Pal AWE system as a comparison to the full W-Pal system because the AWE system best represents the type of standard practice common in computer-based writing instruction (i.e., students write an essay, receive feedback, and revise the essay). Thus, in this study, we are comparing the benefits of explicit strategy instruction and targeted strategy practice (via games) combined with essay writing to standard computer-based writing instruction.

2.1 Participants

Participants include 64 high school students from the metro Phoenix area. Students ranged in age from 14 to 19 (M = 15.9, SD = 1.3) and ranged in grade level from 9 to 12 (M = 10.2, SD = 1.0). The students participated in one of two conditions: the W-Pal condition (n = 33) or the AWE condition (n = 31). Twenty-seven of the participants self-identified as English Language Learners (ELLs). The remaining participants self-identified as native speakers of English (NS). In the W-Pal condition, 23 participants self-identified as NSs and 10 self-identified as ELLs. In the AWE condition, 14 participants self-identified as NSs and 17 self-identified as ELLs.

2.2 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 counterbalanced prompts (i.e., the value of competition/cooperation; the effects of images/impressions). Sessions 2-9 were devoted to training. The students in the W-Pal condition used the full W-Pal. The students in the AWE condition interacted only with the essay writing and automated feedback tools in W-Pal. Thus, a major contrast between the two groups is the number of essays written. Participants in the W-Pal group wrote and received feedback on 8 essays, whereas students in the AWE condition wrote and received feedback on 16 essays (i.e., more essay practice). Time on task in the two conditions was equivalent.

2.3 Corpus and Scoring

The final corpus of essays used in this analysis comprised 128 pretest and posttest essays written by the 64 participants. Descriptive corpus statistics are presented in

Table 1. The essays were scored using the automated scoring algorithm implemented within the W-Pal AWE system. The scoring algorithm assesses essay quality using a combination of computational linguistics and statistical modeling as discussed in [20]. Briefly, the algorithm initially partitions essays into low and high proficiency bins based on number of words and paragraphs thresholds. In subsequent stages, the model presumes that essays that meet and do not meet these thresholds can be characterized by different linguistic features related to lexical sophistication, syntactic complexity, cohesion, semantic categories, and rhetorical elements. Following the initial partition, a number of machine learning algorithms are calculated separately for each group. Each of these algorithms are assigned low proficiency essays a score of 1, 2, or 3 and high proficiency essays a score of 3, 4, 5, or 6.

Table 1. Descriptive statistics for essay corpus: *M* (*SD*)

Paragraphs	Sentences	Words
3.594 (1.359)	21.016 (8.444)	387.211 (129.932)

2.4 Selected Cohesion Indices

We selected a number of local-level cohesion indices (i.e., argument overlap, verb overlap, incidence of *and,* and incidence of all connectives) and global-level cohesion indices (i.e., givenness and incidence of conjuncts) from Coh-Metrix. We also selected newly developed automated indices of global cohesion from the WAT that were created specifically for assessing writing quality. These indices assess cohesion at the paragraph level.

Argument Overlap. Argument overlap refers to the extent to which arguments (nouns, pronouns, and noun phrases) overlap between sentences. Coh-Metrix measures argument overlap between adjacent sentences.

Verb Cohesion. The WAT calculates verb overlap using LSA by computing the average cosine between verbs in adjacent sentences. This index is indicative of the extent to which verbs are repeated across sentences.

Givenness. Given information is information that is recoverable from the preceding discourse. Coh-Metrix calculates text givenness using perpendicular and parallel Latent Semantic Analysis (LSA) vectors [21]. Givenness is computed across a text.

Connectives. Connectives make the relationships among clauses and sentences more explicit. Coh-Metrix assesses negative, positive, additive, temporal, and causal connectives along with conjuncts. These indices are combined into an overall count of connectives. We also include two individual connective scores: incidence of *and* and incidence of conjuncts (e.g., *however* and *in addition*).

Paragraph Cohesion. The WAT measures paragraph cohesion by computing semantic overlap between paragraph types (initial to middle, middle to final, and initial to final). These indices use LSA vectors to compare paragraph types.

2.5 Statistical Analysis

To assess potential differences in prior writing proficiency between NS and ELL participants and between the randomly assigned W-Pal and AWE conditions, we first conducted *t*-tests to compare the automated essay scores at pretest. We also compared scores for the two prompts to ensure that prompt-based effects did not exist. Finally, to assess differences between the pretest and posttest essays for each condition, we conducted mixed-factor analyses of variance (ANOVA) for the selected cohesion indices. We included condition (W-Pal or AWE) as a between-subjects factor.

3 Results

3.1 Differences between NSs and ELL Participants

There was no statistical difference in writing quality as measured by the scoring algorithm between ELL (M = 2.593, SD = .931) and NS participants (M = 2.351, SD = .887), (t = 1.051, df = 62, p = .297). This finding indicates that the NS and ELL participants were of equal writing proficiency at the pretest.

3.2 Differences between Conditions

There was no statistical difference in pretest writing quality for the participants in the W-Pal (M = 2.488, SD = 1.064) and the AWE condition (M = 2.419, SD = .721), (t = .286, df = 62, p = .775). This finding indicates that the writers in both conditions were of equal writing proficiency at the pretest.

3.3 Differences between Prompts

There was no statistical difference between the writing prompts *Images* (M = 2.778, SD = .906) and *Competition* (M = 2.635, SD = 1.222) for all the essays in the corpus, (t = .894, df = 62, p = .375). This finding indicates that there were no prompt-based writing effects for the assigned scores.

3.4 Repeated-Measures ANOVAs for Cohesion Features

There was a significant main effect of test for the following cohesion features: incidence of conjuncts, incidence of *and*s, LSA givenness, LSA middle to middle paragraphs, and LSA middle to final paragraphs. No significant effects were reported for connectives, argument overlap, verb overlap, LSA initial to middle paragraph, and LSA initial to final paragraph (see Table 2 for ANOVA results). These results indicate that participants produced essays that exhibited increased local and global cohesion in the posttest as compared to the pretest (see Table 1 for mean scores in the pretest and posttest). No linguistic features showed a significant interaction between test and condition. These results indicate that the two modes of instruction and practice were equally effective for developing cohesion.

Table 2. Mean (*SD*) and *F* for cohesion indices

Local indices	Pretest	Posttest	*F*
Ands	0.987 (0.557)	1.232 (0.855)	5.147*
All connectives	96.961 (19.894)	98.145 (17.872)	0.199
Argument overlap	0.533 (0.179)	0.497 (0.184)	2.410
Verb overlap	0.107 (0.039)	0.113 (0.035)	1.396

Global indices	Pretest	Posttest	*F*
Conjuncts	0.344 (0.287)	0.519 (0.369)	12.513**
LSA givenness	0.313 (0.043)	0.336 (0.046)	12.292**
LSA I-to-M	0.051 (0.245)	0.166 (0.431)	2.879
LSA I-to-F	0.124 (0.311)	0.196 (0.029)	1.829
LSA M-to-M	0.090 (0.436)	0.281 (0.519)	5.257*
LSA M-to-F	0.097 (0.422)	0.309 (0.605)	4.742*

Note: I = initial paragraph, M = middle paragraph, F = final paragraph
* $p < .050$, ** $p < .001$

4 Discussion

We present an evaluation of the W-Pal ITS through the use of computational indices related to text cohesion. This study demonstrates that automated indices of text cohesion can be used to assess the effects of writing instruction. For both the ITS and the AWE systems, student interaction led to increased use of cohesion features in essay writing. Thus, the use of both the W-Pal ITS and the W-Pal AWE systems can promote writing development, at least with respect to certain cohesive devices.

The students who took part in the W-Pal and the AWE condition demonstrated growth in a variety of cohesion features, including the use of conjuncts, the use of *and*, the increase in given information, and greater semantic overlap between middle paragraphs, and middle and final paragraphs. These findings demonstrate that a mixture of writing instruction, game play, and automated feedback as found in the W-Pal condition led to an increased use of some cohesion features from the pretest to the posttest writing samples. These findings also indicate that intensive writing practice coupled with automated feedback, as found in the AWE condition, also leads to greater production of some cohesion features.

Overall, we found no differences in cohesion scores between the two conditions even though the students in W-Pal condition wrote and revised half as many essays as the essay writing condition. Thus, students who received a mix of writing instruction, practice games, and essay practice with feedback showed similar gains in automated cohesion scores as students who only wrote and revised essays with feedback. Studies have demonstrated that essay-based practice is effective in training writers to increase writing skills [15-16]. However, such practice may be highly repetitive and lower student motivation [20]. The findings from this study suggest that a successful alternative to repetitive essay-based practice is the use of a writing ITS such as W-Pal.

Unlike an AWE system, an ITS provides students not only with the opportunity to practice writing and receive feedback, but also with opportunities to learn writing strategies and play educational games. This mix of options appears to lead to similar gains in cohesion scores as repetitive essay-based practice alone.

The automated cohesion features that demonstrated development over the course of the study are generally related to *global cohesion*. Thus, students in W-Pal and the W-Pal AWE system seemed to develop more global elements of text organization (excluding the increase in the use of *and*) making connections between larger text segments. For instance, conjuncts can not only be used to connect sentences, but also paragraphs. Conjuncts can also be used to provide global organization through enumeration (i.e., *first, second, third*) and summarizing (*to sum up*). Givenness provides information about the use of new and old information across a text. Lastly, our paragraph cohesion indices measure semantic similarity at the global level. Previous research [12] has reported correlations between global cohesion indices and human judgments of text coherence. Such findings along with those reported here suggest that writers working within the W-Pal ITS and AWE systems may begin to develop texts that are more globally coherent. Since indices of global coherence are also linked to essay quality [12], their use may lead to better quality essays.

The majority of the indices that did not demonstrate significant change from pretest to posttest measured *local cohesion* (e.g., general connectives and argument and verb overlap between adjacent sentences). This finding suggests that writers using W-Pal or the W-Pal AWE system do not focus on developing connections between smaller elements of text (i.e., local cohesion). The exceptions were the paragraph cohesion measures that involve the initial paragraphs. Initial paragraphs generally include many textual functions such as an introduction, a claim, and arguments. Thus, initial paragraphs may not overlap strongly with body and conclusion paragraphs because of the number and variety of the textual functions they contain. However, body paragraphs should be semantically related in that they develop similar themes. In addition, conclusion paragraphs should demonstrate greater semantic overlap with body paragraphs because they should include a summary of the body paragraphs.

In general, these findings support earlier research, which has suggested that indices of local cohesion were not significant predictors of essay quality [10], but that indices of global cohesion were [11]. Thus, as writers develop and essay quality increases, we should expect to see a greater development and use of global cohesion in essays, but not in local cohesion.

5 Conclusion

Overall, this study demonstrates how computational indices of cohesion can be used to evaluate ITS and AWE systems. In addition, this study demonstrates how such indices can be used to assess student writing in terms of the development and use of local and global cohesion in essays. Such evaluations can help explain the efficacy of ITSs as compared to AWE systems and help to examine writing development in adolescent learners. In this study, we find that ITS systems are as effective as AWE

systems in terms of the development of cohesion strategies even when users of the AWE systems write twice as many essays. We also find that the majority of global cohesion indices show gains between pretest and posttest writing whereas the majority of local cohesion indices do not.

While these findings suggest positive effects of both the W-Pal and the AWE system on writing, additional studies are needed to demonstrate equivalence between the two approaches. Such studies will require a comprehensive investigation of all aspects of the two systems and their effects of writing quality, writing development, system engagement, and participant motivation (to name but a few aspects).

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.

References

1. National Commission on Writing: The Neglected "R". College Entrance Examination Board, New York (2003)
2. Kellogg, R., Raulerson, B.: Improving the Writing Skills of College Students. Psychonomic Bulletin and Review 14, 237–242 (2007)
3. Graham, S., Perin, D.: A meta-analysis of writing instruction for adolescent students. Journal of Educational Psychology 99, 445–476 (2007)
4. McGarrell, H., Verbeem, J.: Motivating revision of drafts through formative feedback. ELT Journal 61, 228–236 (2007)
5. Devillez, R.: Writing: Step by Step. Kendall Hunt, Dubuque (2003)
6. Golightly, K., Sanders, G.: Writing and Reading in the Disciplines, vol. 2. Pearson, New Jersey (2000)
7. McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G.T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., Graesser, A.: The Writing-Pal: Natural Language Algorithms to Support Intelligent Tutoring on Writing Strategies. In: McCarthy, P.M., Boonthum-Denecke, C. (eds.) Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution, pp. 298–311. IGI Global, Hershey (2012)
8. Graesser, A., McNamara, D., Louwerse, M., Cai, Z.: Coh-Metrix: Analysis of Text on Cohesion and Language. Behavioral Research Methods, Instruments and Computers 36, 193–202 (2004)
9. McNamara, D., Crossley, S., Roscoe, R.: Natural Language Processing in an Intelligent Writing Strategy Tutoring System. Behavioral Research Methods, Instruments and Computers (2012) (Advance online publication)
10. McNamara, D., Kintsch, E., Songer, N., Kintsch, W.: Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. Cognition and Instruction 14, 1–43 (1996)
11. Crossley, S., McNamara, D.: Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. In: Ohlsson, S., Catrambone, R. (eds.) Proceedings of the 32nd Annual Conference of the Cognitive Science Society, pp. 984–989. Cognitive Science Society, Austin (2010)

12. Crossley, S., McNamara, D.: Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. In: Carlson, L., Hoelscher, C., Shipley, T.F. (eds.) Proceedings of the 29th Annual Conference of the Cognitive Science Society, pp. 1236–1241. Cognitive Science Society, Austin (2011)
13. Crossley, S., Weston, J., Sullivan, S., McNamara, D.: The Development of Writing Proficiency as a Function of Grade Level: A Linguistic Analysis. Written Communication 28, 282–311 (2011)
14. Crossley, S., McNamara, D.: Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication. Journal of Research in Reading 53, 115–136 (2012)
15. Johnstone, K.M., Ashbaugh, H., Warfield, T.D.: Effects of repeated practice and contextual-writing experiences on college students' writing skills. Journal of Educational Psychology 94, 305–315 (2002)
16. Kellogg, R.T., Raulerson, B.A.: Improving the writing skills of college students. Psychonomic Bulletin & Review 13(2), 237–242 (2007)
17. Graham, S., Harris, K.R.: The role of self-regulation and transcription skills in writing and writing development. Educational Psychologist 35, 3–12 (2000)
18. Grimes, D., Warschauer, W.: Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. Journal of Technology, Learning, and Assessment 8, 4–43 (2010)
19. Roscoe, R., Kugler, D., Crossley, S., Weston, J., McNamara, D.: Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In: McCarthy, P., Youngblood, Y. (eds.) Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, pp. 466–471. The AAAI Press, Menlo Park (2012)
20. Crossley, S., Roscoe, R., McNamara, D.: Using Natural Language Processing Algorithms to Detect Changes in Student Writing in an Intelligent Tutoring System. Manuscript Submitted to the 26th International Florida Artificial Intelligence Research Society Conference (2013)
21. Hempelmann, C., Dufty, D., McCarthy, P., Graesser, A., Cai, Z., McNamara, D.: Using LSA to Automatically Identify Givenness and Newness of Noun Phrases in Written Discourse. In: Bara, B.G., Bucciarelli, M. (eds.) Proceedings of the 27th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Mahwah (2005)