

Exploring Individual Differences in Student Writing with a Narrative Composition Support Environment

Julius Goth and Alok Baikadi and Eun Ha and Jonathan Rowe and Bradford Mott
and James Lester

Department of Computer Science
North Carolina State University
Raleigh, NC, USA

{jgoth, abaikad, eha, jprowe, bwmott, lester}@ncsu.edu

Abstract

Novice writers face significant challenges as they learn to master the broad range of skills that contribute to composition. Novice and expert writers differ considerably, and devising effective composition support tools for novice writers requires a clear understanding of the process and products of writing. This paper reports on a study conducted with more than one hundred middle grade students interacting with a narrative composition support environment. The texts are found to pose important challenges for state-of-the-art natural language processing techniques. Furthermore, the study investigates the language usage of middle grade students, the cohesion and coherence of the resulting texts, and the relationship between students' language arts skills and their writing processes. The findings suggest that composition support environments require robust NLP tools that can account for the variations in students' writing in order to effectively support each phase of the writing process.

1 Introduction

Writing is fundamentally complex. Writers must simultaneously consider a constellation of factors during composition, including writing task requirements, knowledge of audience, domain knowledge, language usage, and tone (Hayes and Flower, 1981). Furthermore, effective writing involves sophisticated higher-order cognitive skills, such as synthesis of ideas, critical thinking,

and self-regulation. Text genres, such as narrative or expository texts, also introduce distinct requirements and conventions (Hayes and Flower, 1981).

Because writing itself is complex, learning to write poses significant challenges for students. The central role of writing in communication, knowledge organization, and sensemaking points to the need to devise methods and tools with which writing skills can be effectively taught and learned (Graham, 2006). Intelligent tutoring systems (VanLehn, 2006) offer a promising means for delivering tailored writing support to students. However, developing intelligent tutors to scaffold student writing poses a number of technical and pedagogical hurdles. Texts written by novice writers are likely to exhibit significant variation in grammar, cohesion, coherence, and content quality; these characteristics are likely to be problematic for analysis by current natural language processing tools. Furthermore, students' individual differences in language arts skills, writing self-efficacy, domain knowledge, and motivation can have pedagogical implications. An effective intelligent writing tutor must do more than just parse and understand student texts; it must also provide tailored feedback that fosters effective writing processes and enhances student motivation for writing.

This paper explores several key questions for the design of intelligent composition support tools for novice writers. First, it investigates the performance of current syntactic parsing tools on a corpus of narrative texts written by middle grade students during interactions in a narrative composi-

tion support environment. A *narrative composition support environment* aims to support the principal processes of writing, such as planning, revision, and text production. The second question the paper explores is how middle school students' language art skills affect the cohesion and coherence of texts produced during interactions with a narrative composition support environment. Third, the paper investigates how middle school students' language art skills affect their writing processes during interactions in a narrative composition support environment. Studying the interactions between the environment's support mechanisms and students' individual differences provides insights into the affordances and limitations of novices' writing abilities, as well as implications for the design of intelligent tutors for narrative writing.

The study presented here investigates novice writers' composition processes during interactions with a narrative composition support environment. In the study, 127 middle grade students interacted with the NARRATIVE THEATRE fable composition support environment. The NARRATIVE THEATRE uses a multimedia interface to guide students as they select key elements of their fable (e.g., moral, setting, characters), prompts students through an explicit, timed story planning process, and allows students to review their earlier planning decisions at any point during writing of the main text. Students' literacy ratings and log data from interactions with the NARRATIVE THEATRE environment are analyzed to investigate the differences between high- and low-skill students and their practice of key composition processes in the NARRATIVE THEATRE environment, including planning, text production, and revision. Coh-Metrix (Graesser et al., 2004) was also used to analyze the cohesion and coherence characteristics of the students' fables. The observations from this study offer important implications for the design of intelligent composition support tools for novice writers.

2 Related Work

Since Hayes and Flower first proposed their seminal model of writing nearly thirty years ago (1981), a rich body of work has investigated the cognitive functions supporting written composition. Foundational results are now in place on the core processes of writing, including idea generation (Galbraith et al., 2009), text production (Berninger

et al., 2002), and revision (McCutchen et al., 1997). Furthermore, a detailed account of the composition process has begun to emerge across a range of writing experience levels (Graham et al., 2002) and text genres (Langer, 1985).

Particularly important for the design of composition support tools for novices is the emergence of a consensus account of the characteristics of novice writers' narrative composition processes. Empirical studies have suggested that notable differences exist between novice and expert writers, such as novices' use of knowledge-telling practices versus experts' use of knowledge-transformation practices during text production (Bereiter and Scardamalia, 1987). However, it has been argued that even novice writers can employ high-level knowledge-transformation processes when situated within an appropriate task environment with effective writing scaffolds (Cameron and Moshenko, 1996). Other work has found that students' domain and linguistic knowledge influences the coherence and quality of their expository writings (DeGroff, 1987). These findings underscore the importance of investigating methods for effective and engaging writing instruction targeted at novice writers, as well as automated tools to tailor feedback and scaffolding to individual students.

In addition to grounding their work in the writing research literature, designers of composition support tools will likely need to avail themselves of the full gamut of natural language processing techniques to analyze students' texts with regard to syntax, semantics, and discourse. However, in texts produced by novice writers, grammatical errors and incoherent discourse abound, which may present serious challenges for natural language processing since the majority of current NLP tools have been developed for well-formed texts. While existing NLP tools have been successfully used in writing support systems designed for expert writers (Mahlow and Piotrowski, 2009), common structural issues in novice compositions are likely to prove problematic for current tools. However, recent work has begun to explore techniques for handling ill-formed texts that are similar to those produced by novice writers. For example, Gamon et al. conducted a word-level analysis of texts written by non-native English speakers (2008). Focusing on two types of errors (determiners and prepositions), they use decision-tree



Figure 1. Narrative Theatre fable composition support environment.

classifiers in combination with a language model trained on a large English corpus to detect and correct erroneous selection of words. Wagner et al. investigated the detection of grammatical malformedness of individual sentences (2007). They found it effective to combine a shallow approach that uses n -grams and a deep approach that uses syntactic parse results. Higgins et al. explored the overall coherence of texts written by students (2004). Using support vector machines, their system identified the portions of text that resulted in coherence breakdowns with regard to relatedness to the essay question and relatedness between discourse elements.

To date, a relatively small number of intelligent tutoring systems have been developed to support student learning in the language arts, and even fewer have sought to specifically address writing. Sourcer's Apprentice is a web-based learning environment to help high school students gather, evaluate, and integrate information for writing essays about history topics (Britt et al., 2004), although Sourcer's Apprentice did not seek to apply NLP tools to understand or scaffold students' compositions directly. Other work on intelligent tutoring

for language arts, such as Project LISTEN (Mostow and Aist, 2001) and REAP (Heilman et al., 2007), has addressed vocabulary learning and reading comprehension.

3 Narrative Corpus Acquisition

To investigate narrative composition in novice writers, a study was conducted with more than one hundred middle grade students using a narrative composition support environment. The NARRATIVE THEATRE (Figure 1) is an interactive environment designed to capture both the process and products of writing.¹ Targeting a user population of sixth grade students (age typically 12 years) and the genre of fables, the NARRATIVE THEATRE enables students to create stories in an environment that was specifically designed to scaffold novices' composition activities during a timed story plan-

¹ The version of the NARRATIVE THEATRE used in the study reported in this paper is the forerunner of a more general creativity support environment. It is under development in our laboratory that will employ NLP techniques and intelligent graphics generation. The study reported here was conducted to inform the design of the creativity enhancement environment and intelligent tutoring systems to support composition.

ning and writing process. The NARRATIVE THEATRE employs a multimedia interface created with Adobe's Flash® development platform and AIR runtime environment. Its design was inspired by a worksheet that is widely used as part of the Grade 6 writing curriculum.

During the planning phase, students select a moral, a setting, a cast of characters, and a set of objects for the story they will create. The system provides nine different morals, four settings, ten characters, and twenty objects from which students may choose. Each setting is accompanied by a visual representation, which can be enlarged by clicking on the image to highlight salient features of the setting. Characters and objects are also visually represented by static graphics, which were designed to be neutral in gender and expression in order to allow students creative choice when filling narrative roles with the characters.

Once the choices have been made, students are presented with a screen that allows them to view their planning decisions and begin structuring their fable. The planning area allows students to make notes about what they would like to have happen during the beginning, middle, and ending. The top of the page contains windows that display the setting, characters, and objects that were chosen earlier, and that can provide more information via a mouseover. Students craft a plan for the beginning (setting and characters are introduced), middle (conflict and problem), and end (conflict resolution) of their stories. For each of the three major segments of the story, they formulate a textual plan. After the planning information is entered, the students may begin writing (Figure 1). They then create the actual prose, which is entered as raw text. The writing and revision phase are supported with a spell-correction facility. All student activities including interface selections and the text streams from planning and writing are logged and time-stamped.

During the study, a total of 127 sixth-grade middle school students (67 males, 60 females) participated in the study. The students ranged in age from 10 to 13. Approximately 38% of the students were Caucasian, 27% African-American, 17% Hispanic or Latino, 6% Asian, 2% American Indian, and the remaining 10% were of mixed or other descent. Students participated as part of their Language Arts class. The study spanned two days for each student involved. On the first day, the

students were seated at a computer and asked to fill out a pre-experiment questionnaire, which required approximately twenty minutes. On the second day, the students were again assigned to a computer. They were presented with the NARRATIVE THEATRE interface, which asked them to enter a unique identification number. Once correctly entered, the students were presented with a short instructional video that described the features and operation of the interface. They were given fifteen minutes to complete the planning activity, which included choosing a setting, main characters, props, and deciding the beginning, middle, and end of their story. Once planning was completed, or time ran out, the students were given another thirty-five minutes to write their fable. After their fable was completed, the students were asked to complete a post-experiment questionnaire. This survey was also allotted twenty minutes for completion. In total, the study lasted ninety minutes.

4 Findings

Three categories of analyses were performed on the NARRATIVE THEATRE corpus: an analysis of natural language processing tool performance (specifically, an analysis of syntactic parsers), an analysis of coherence and cohesion in the written texts using the automated cohesion metric tool Coh-Metrix (Graesser et al., 2004), and an analysis of students' writing processes.

As part of an investigation of students' individual differences in writing, students' language arts skills were measured by their scores from the prior year's End-of-Grade reading test. Subjective ratings of writing ability were also obtained for each student from their teachers. The reading scores were used in the presented analyses because they were obtained through systematic testing, but it is interesting to note that the objective reading scores and subjective writing scores were found to be strongly correlated by calculating the Spearman's correlation coefficient², $\rho = .798$, $p < .0001$. The high correlation suggests that reading scores can serve as a reasonable indicator of language arts skills.

² Spearman's correlation coefficient was used because of the ordinal nature of the reading and writing measures (Myers and Well, 2003).

Coh-Matrix feature	Below-Grade	At-Grade	Above-Grade	F(2, 110) =
Hypernym, nouns	5.9 (0.93)	6.17 (0.81)	6.53 (0.43)	1.24 Below-Above**
Hypernym, verbs	1.44 (0.18)	1.49 (0.18)	1.48 (0.17)	3.72
Causal cohesion	0.83 (0.09)	0.87 (0.1)	0.39 (0.16)	3.70 Below-Above** At-Above**
LSA, paragraph to paragraph	0.34 (0.19)	0.45 (0.22)	0.49 (0.18)	3.89 Below-Above** Below-At*
LSA, sentence to sentence	0.21 (0.14)	0.24 (0.11)	0.22 (0.09)	0.48
Personal pronoun usage	107 (35.88)	101 (29.98)	89 (19.37)	2.25 Below-Above*
Pronoun to noun phrase ratio	0.36 (0.12)	0.35 (0.10)	0.30 (0.06)	2.21

Table 1. The effects of reading grade-level on select Coh-Matrix features.

* denotes $p < .1$ and ** denotes $p < .05$

4.1 Natural Language Processing

Two syntactic parsing tools were used to analyze students' fables and develop an initial account of the performance of current natural language processing tools on a corpus of novice-generated narrative texts. The Link Grammar Parser (Temperley, 1995) and Stanford Parser (Klein and Manning, 2003) were run on the entire corpus, and their performance recorded.

Link parsing provides insight into the number of grammatically malformed sentences observed in each fable. Link grammars center on the notion of linkable entities directly combined with one another, as opposed to tree-like structures. Link parsers attempt to identify one or more syntactically valid representations, where each entity is paired with another. Passages were split into sentences using OpenNLP, and then run against the Link Grammar Parser (Temperley, 1995). If a sentence had no suitable link based on the parser (e.g., "Last dog I saw a great movie"), it was considered "broken" because it lacked an appropriate linkage. A ratio of sentences without appropriate linkage to total sentence count was used to characterize the link parser's performance on each student's fable.

On average, the Link Grammar Parser found linkages for 41% of sentences ($SD=.22$). Interestingly, reading level was shown to have a marginal effect on the link parser's success rate, $F(2,110) = 5.78$, $p = .06$. Post hoc Tukey's tests revealed that above-grade level readers were marginally more likely to write linkable sentences than at-grade level readers, $p = .07$. The effect was

strongly significant between above-grade level readers and below-grade level readers, $p = .003$.

The Stanford parser was used to investigate the frequency with which sentences could be successfully parsed. A parsing failure was noted any time the tool was forced to fall back to a PCFG parse. On average, the Stanford parser produced a parse for 91% of students' sentences. A significant effect of reading grade-level on Stanford parser success rate was observed, $F(2,110) = 4.41$, $p = .015$. Post hoc tests showed that above-grade level readers wrote significantly more sentences that could be parsed than below-grade level readers, $p = .02$. There was also a marginal difference observed between below-grade level readers and at-grade level readers, $p = .08$.

Gender was not found to have an effect on the percentage of linkable sentences, nor the number of Stanford parser failures.

4.2 Individual Differences and Written Texts

Several analyses were conducted to investigate individual differences in students' written texts. Analyses focused on writing length, cohesion characteristics, coherence characteristics, and spelling errors. Fable lengths were measured in characters ($M = 1346$, $SD = 601$).

A marginal effect of reading grade-level on fable length was observed, $F(2,110) = 2.89$, $p = .06$. Post hoc tests showed that at-grade level readers tended to write longer fables than below-grade level readers, $p = .10$. Gender was also found to have a significant effect on writing length. Specifically, females tended to write longer fables than males, $F(1,110) = 4.41$, $p = .04$.

Writing process feature	Below-Grade	At-Grade	Above-Grade	F(2, 110) =
Avg length of deletion, planning	21.30 (8.31)	28.18 (11.14)	30.99 (12.37)	8.34 Below-Above*** Below-At***
Avg length of deletion, writing	23.42 (9.26)	27.74 (10.28)	33.06 (16.80)	5.18 Below-Above***
Mouseovers/min	0.19 (0.15)	0.09 (0.07)	0.07 (0.05)	11.81 Below-Above*** Below-At***
5+ second revision count, planning	9.14 (6.62)	5.43 (4.43)	2.37 (2.36)	12.70 Below-Above*** Below-At*** At-Above*
5+ second revision count, writing	18.04 (7.84)	14.21 (7.81)	13.37 (7.52)	3.83 Below-Above* Below-At*

Table 2. The effects of reading grade-level on writing process characteristics.

* denotes $p < .1$, ** denotes $p < .05$, and *** denotes $p < .01$.

To investigate cohesion and coherence in students' fables, the corpus was analyzed with Coh-Metrix, a tool for analyzing the cohesion, language, and readability of texts (Graesser et al., 2004). At the core of Coh-Metrix is a lexical analyzer, syntactic parser, part-of-speech tagger, and reference corpora (for LSA) that processes text and returns linguistic and discourse related features. Coh-Metrix measures several types of cohesion, as well as concreteness, connectives, diversity in language, and syntactic complexity. Concreteness is measured using the hypernym depth values retrieved from the WordNet lexical taxonomy, and averaged across noun and verb categories.

Results from an analysis of reading grade-levels and Coh-Metrix features are presented in Table 1. Interestingly, above-grade level students were observed to have lower causal cohesion scores than at-grade level or below-grade level students. The converse is found in an examination of paragraph-to-paragraph LSA scores, which are often used to measure semantic cohesion. Below-grade level readers tended to have lower semantic cohesion scores than at-grade level readers. LSA scores on adjacent sentences and all combinations of sentences were not significant across any of the groups. Sentence-to-sentence LSA scores were also not significant across groups.

Gender did not have a significant effect on causal cohesion, hypernym depth of verbs, or paragraph-to-paragraph LSA values. However, gender was found to have a significant effect on hypernym depth of nouns, $F(1,110) = 15.96$, $p = .0001$. Males tended to use more concrete nouns in their writing passages, with an average difference of .6 in hypernym depth. The ratio of

pronouns to noun phrases was also significant between genders, $F(1,110) = 10.19$, $p = .002$. Females had a 38% pronoun to NP ratio whereas males were at 32%. Gender had a significant effect on sentence-to-sentence LSA scores, $F(1,110) = 19.9$, $p = .0001$. Males tended to have a higher LSA score across adjacent sentences ($M = .27$, $SD = .11$) than females ($M = .18$, $SD = .1$). Finally, gender had a significant effect on personal pronoun incidence score, $F(1,110) = 9.12$, $p = .003$. Females used personal pronouns as 11.1% of their content whereas males used them as 9.3% of their content.

An examination of the number of spelling errors remaining in student fables, as well as students' usage of the built-in spelling corrector, was conducted. However, no significant effects were observed across reading level or gender.

4.3 Individual Differences and Writing Processes

Several features in the student interaction logs were chosen to investigate key aspects of students' writing processes. Specifically, these features include planning length, planning and writing time, revision behavior, pauses in text production, and reviews of prior planning decisions.

On average, students spent 665 seconds planning their fables and 2199 seconds writing their fables ($SD = 535$). Students also typed 537 characters on average while planning their fables ($SD = 254$). No significant effect of reading level was observed on planning length, but reading level did have a significant effect on time spent in the planning phase, $F(2,110) = 12.76$, $p < .0001$. Below-grade level readers spent significantly more time

on planning than at-grade level readers, $p = .001$, as well as above-grade level readers, $p < .0001$. There were also significant differences in writing time across reading level groups, $F(2, 110) = 6.47$, $p = .002$. Below-grade level readers took significantly more time composing their fables than at-grade level readers, $p = .05$. Also, below-grade level readers took significantly more time to write their fables than above-grade level readers, $p = .003$.

Females tended to write longer passages in the planning section than males $F(1,110) = 4.68$, $p = .03$. Time spent on the planning section was lower among females than males, $F(1,110) = 3.92$, $p = .05$. Females also spent less time on the writing section than males, $F(1,110) = 3.87$, $p = .05$.

Students' revision behaviors were gauged using a heuristic that measures edit distances between successive snapshots of fables collected at one-minute intervals during composition. Each minute, a static snapshot of student's fable progress was taken and logged. Edit distances between successive snapshots of students' fables were measured using the Google Diff, Match and Patch tools to make "before" and "after" comparisons (Google, 2009). Comparing two successive snapshots of a single fable, a revision was defined as any insertion of text that occurred before the tail end of the fable.

The effects of reading level on revision in both the planning and writing stages is presented in Table 2. During the writing stage, a significant effect of grade-level was observed on average revision length between below-grade level readers and at-grade level readers, as well as between below-grade level readers and above-grade level readers. Within the planning section, at-grade level readers revised more text than below-grade level readers, and above-grade level readers revised more text than at-grade level readers. Self-efficacy for writing was found to be significantly correlated with average revision length in both the planning, $r = .21$, $p = .03$, and writing, $r = .31$, $p = .001$, stages.

Pauses between successive keystrokes were investigated during both the planning and writing stages of NARRATIVE THEATRE interactions. For the purpose of this work, a pause is defined as a keystroke made five or more seconds after the preceding keystroke. Keystroke pauses were categorized as either an appendage or a revision, depending on whether they occurred before the tail

end of the passage (revision) or after the tail end (appendage). For the planning section, below-grade readers paused significantly more often than at-grade readers. Also, at-grade readers paused before revising significantly more often than above-grade readers. The effects of reading level on a number of writing process subscores are shown in Table 2.

Gender had a significant effect on pauses prior to revision in the writing phase, $F(1,110) = 3.26$, $p = .07$. Females paused on more occasions than males. However, no gender effect was found for pause behavior during the planning phase.

During the planning and writing stages of NARRATIVE THEATRE interactions, students could review their prior planning selections—including characters, objects, and settings—by hovering the mouse over the respective region near the top of the screen (mouseover). Upon hovering the mouse over the appropriate region, a graphical illustration of the student's planning selection was presented. Mouseover instances were recorded to obtain insight into idea generation, or instances where the student was contemplating what to write next. Mouseovers were calculated in terms of average mouseovers over time (in minutes). The effects of reading ability on mouseover behaviors are shown in Table 2.

For the mouseover metric, reading level had a significant effect on the mouseover rate. Below-grade level learners tended to use the mouseover feature on a more frequent basis than both at-grade level and above-grade level readers. There was not a significant difference between at-grade level and above-grade level groups.

The effect of gender on mouseover rate was significant, $F(1,110) = 9.93$, $p = .002$. Males used the mouseover feature on fewer occasions than females.

5 Discussion

The performance of the two parsers differed widely. The Stanford Parser was able to parse over 90% of fables, but the Link Grammar Parser was only successful for about 40% of the fables. While parser failure is not always indicative of poor grammaticality, every sentence that failed on the Stanford Parser contained either misspelled words or run-on sentences. Many of these were indicated by errors in the sentence segmentation as well.

There were also indications that students' language arts skills may influence the grammaticality of their written sentences; significant effects of reading level were found on both the Stanford Parser's success rate and the Link parser's success rate. The fact that below-grade level students constituted a considerable proportion of the study's student population suggests that pedagogical writing support tools should be capable of handling the variations inherent in students' writings, and leverage natural language processing results to inform tutorial feedback.

Paragraph-to-paragraph LSA scores tended to increase with reading level. This has implications for semantic cohesion (Graessar, 2004) and indicates that students with a higher reading assessment score produce stories that satisfy this particular dimension of cohesion. However, the converse was true for the Coh-Metrix measure of causal cohesion, where above-grade level students actually produced the lowest cohesion scores. One possible explanation could stem from differences in vocabulary skills between above- and below-grade level students; students who exercise a larger vocabulary may be penalized by Coh-Metrix's cohesion metric. Alternatively, the result may be related to the fact that below-grade level students tend to produce less text (Graessar, 2004). Clearly, students' individual differences in language arts ability affect the cohesiveness of the texts they write, but additional investigation is necessary to develop a clear understanding of the relationship between cohesion and language arts ability, as well as the implications for tailoring tutoring.

With regard to the writing process, the average length per revision was significantly greater for students of higher reading skill-levels. There is a possibility that this may be associated with more elaborate revision processes, which requires further investigation. It should be noted that the revision finding was more salient for the planning stage of NARRATIVE THEATRE interactions. This result may also indicate that below-grade level readers were somewhat less thorough when planning their fables. Further, differences in mouseover behavior were found across reading levels, apparently indicating a decline in the rate of mouseovers as reading level increased. This finding may be the result of below-grade level students experiencing difficulties in idea generation, or a lack of motivation. Finally, the number of pauses prior to revision was

found to decrease as reading level increased. This result may point to difficulties with text production for lower language arts skill students. Difficulty translating ideas into text may point to a need for intelligent writing tutors to help reduce lower reading level students' cognitive load during writing.

6 Conclusions and Future Work

We have presented a study conducted with middle grade students to investigate the process and products of writing in a narrative composition support environment. The study found significant variations in syntactic parser performance associated with students' language arts abilities, as well as relationships between students' reading level and the grammaticality of their writing. For example, the stories of below-grade readers had a lower level of semantic cohesion than at-grade level readers, but surprisingly, above-grade level students' writings exhibited lower causal cohesion than both at-grade and below-grade level students. Reading level had a significant effect on time spent in the planning phase, and below-grade level readers spent more time composing fables than at-grade level readers. There were also gender differences, with females spending less time in both the planning and writing phases. There were also differences with respect to revision, with above-grade readers revising more than below-grade readers.

The study highlights important issues about how to design composition support tools. Composition support tools that are sensitive to students' individual writing abilities seem likely to be most effective. Natural language processing is critical for analyzing students' texts and informing the content of adaptive tutorial feedback. Intelligent writing tutors should utilize natural language processing techniques that can robustly handle the variations in students' writings, and deliver tailored scaffolding informed by analyses of students' texts and writing processes.

The findings suggest that several directions exist for future work. Additional analysis is necessary to investigate the correctness of syntactic parses. Further investigation of students' individual differences in writing at the discourse and narrative levels is also necessary. Results from these analyses should then be used to inform the design of techniques for adaptive tutorial feedback in narrative composition support environments.

Acknowledgements

The authors wish to thank members of the North Carolina State University IntelliMedia Group for their assistance with the NARRATIVE THEATRE. This research was supported by the National Science Foundation under Grant IIS-0757535. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- C. Bereiter and M. Scardamalia. 1987. *The Psychology of Written Composition*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- V. W. Berninger, K. Vaughan, R. D. Abbott, K. Begay, K. B. Coleman, G. Curtain, J. M. Hawkins, and S. Graham. 2002. Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology*, 94(2):291–304.
- M. A. Britt, P. Wiemer-Hasting, A. Larson, and C. Perfetti. 2004. Automated feedback on source citation in essay writing. *International Journal of Artificial Intelligence in Education*, 14(3–4):359–374.
- C. A. Cameron and B. Moshenko. 1996. Elicitation of knowledge transformational reports while children write narratives. *Canadian Journal of Behavioural Science*, 28(4):271–280.
- L. J. C. DeGross. 1987. The influence of prior knowledge on writing, conferencing, and revising. *Elementary School Journal*, 88(2):105–118.
- L. Flower and J. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.
- D. Galbraith, J. Hallam, T. Olive, and N. Le Bigot. 2009. The role of different components of working memory in writing. In *Proceedings of Annual Conference of the Cognitive Science Society*, Amsterdam, The Netherlands.
- M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 449–456, Hyderabad, India.
- Google Diff Match and Patch [Software]. Available from <http://code.google.com/p/google-diff-match-patch/>
- A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers*, 36(2):193–202.
- S. Graham. 2006. Strategy instruction and the teaching of writing: A meta-analysis. In C. A. MacArthur, S. Graham, and J. Fitzgerald, editors, *Handbook of writing research*. Guilford Press, New York, NY, pages 187–207.
- S. Graham, K. R. Harris, and B. F. Chorzempa. 2002. Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology*, 94(4):669–686.
- J. Hayes. 1996. A new framework for understanding cognition and affect in writing. In C. M. Levy and S. Ransdell, editors, *The Science of Writing: Theories, Methods, Individual Differences, and Applications*. Lawrence Erlbaum Associates, Mahwah, NJ, pages 1–28.
- M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of Human Language Technology Conference*, pages 460–467, Rochester, NY.
- D. Higgins, J. Bustein, D. Marcu, and C. Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of Human Language Technology conference/North American chapter of the Association for Computational Linguistics*, pages 185–192, Boston, MA.
- J. A. Langer. 1985. Children's sense of genre: A study of performance on parallel reading and writing Tasks. *Written Communication*, 2(2):157–187.
- C. Mahlow and M. Piotrowski. 2009. LingURed: Language-aware editing functions based on NLP resources. In *Proceedings of International Multiconference on Computer Science and Information Technology*, pages 243–250, Mragowo, Poland.
- D. McCutchen, M. Francis, and S. Kerr. 1997. Revising for meaning: Effects of knowledge and strategy. *Journal of Educational Psychology*, 89(4):667–676.
- J. Mostow and G. Aist. 2001. Evaluating tutors that listen: an overview of project LISTEN. In K. Forbus and P. Feltovich, editors, *Smart Machines in Education*. MIT Press, Cambridge, MA, USA, pages 169–234.
- J. Myers and A. Well. 2003. *Research Design and Statistical Analysis*. Erlbaum, Mahwah, NJ.
- K. VanLehn. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265.
- J. Wagner, J. Foster, and J. Van Genabith. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 112–121, Prague, Czech Republic.