Evaluating the Effects of Natural Language Generation Techniques on Reader Satisfaction

Charles B. Callaway (cbcallaw@eos.ncsu.edu)
James C. Lester (lester@csc.ncsu.edu)
The IntelliMedia Initiative
Department of Computer Science
North Carolina State University
Raleigh, NC 27695 USA

Abstract

We are witnessing the emergence of a new technology for dynamically creating stories tailored to the interests of particular readers. *Narrative prose generators* offer much promise for literacy education, but designing them for maximal effectiveness requires us to understand their effect on readers. This article describes the evaluation of STORYBOOK, an implemented narrative prose generation system that produces original fairy tales in the Little Red Riding Hood domain. STORYBOOK creates two to three pages of text consistently represented at the deep linguistic structure level. Because of this, we can formally evaluate multiple versions of a single story and be assured that the content is identical across all versions. We produced five such versions of two separate stories which were compared by a pool of twenty upper division students in English and analyzed with an ANOVA test. While the results are most informative for designers of narrative prose generators, it provides important baselines for research into natural language systems in general.

Introduction

The emerging technology of narrative prose generation, which dynamically creates stories tailored to the interests of particular readers, offers great promise for literacy education. However, to design effective narrative prose generation software for literacy education, it is important to understand how students perceive texts created by these algorithms. Do the results of studies based on human-produced texts apply? How does computer control of minute aspects of text production affect readers? Do readers have quantitative reactions to fundamental alterations in texts as we expect they would?

As a result of recent work in formally evaluated language generation technology (Smith & Hipp 1994; Robin & McKeown 1995; Allen *et al.* 1996; Callaway & Lester 1997; Lester & Porter 1997; Young 1999), we are seeing an increased awareness of the issues involved in successfully generating texts dynamically for specific target audiences. However, these systems are focused more towards task effectiveness evaluations or explanation generation and are not suitable for the significant difficulties in creating literary narratives. And while there exist story generation systems capable of producing narratives (Meehan 1976; Lebowitz 1985; Lang 1997), none of these systems has been formally evaluated by readers. Furthermore, various formal studies on reading comprehension (Kintsch & Keenan 1973; Graesser, Millis & Zwaan 1997; Hoover 1997) have focused on mechanical aspects such as reading rate, and did not have access to computational mechanisms for producing the texts they studied.

To study the changes in perceived text quality stemming from alterations to the underlying text generation architecture, we conducted a formal study gauging the satisfaction of subjects reading narratives. The study involved the following:

- A consistent representation mechanism which allows for the representation of characters, props, locations, actions and descriptions found in a narrative environment. Holding these entities constant for the duration of an experiment ensures that the stories seen by the study participants will have identical plots and details except for the variations cued from the experiment's parameters.
- A story generation mechanism that, when given the story representation and the experimental parameters, can produce a specified set of narratives. Our story generator, named STORYBOOK, creates narratives in the Little Red Riding Hood fairy tale domain. These narratives can be tailored to produce a variety of grammatical, lexical, and propositional effects.
- A pool of readers familiar with narratives and the writing process itself. Thus we conducted a study involving 20 upper division undergraduate students majoring in English or Communication. Each student read two distinct Little Red Riding Hood stories averaging two hours per student.

There are two primary types of comparisons upon which an evaluation of a text-producing system can focus: human text *vs.* computer text and computer text *vs.* computer text. Although there are a number of pre-existing Little Red Riding Hood texts available for comparison via the World Wide Web, formally comparing such narratives with those produced by computer presents a difficult problem: there is no known objective metric for quantitatively evaluating narrative prose in terms of how it performs *as a story*. Simple metrics exist for evaluation at the sentence level (*e.g.*, number of words, depth of embedding, *etc.*), but a narrative *per se* cannot be considered to be just a collection of sentences that are not related to each other. In addition, because narrative is not a “deductive” domain, it cannot be evaluated in terms of *correctness* by a panel of human judges. To overcome these problems, we instead opted for a computer *vs.* computer style of evaluation that investigates whether certain architectural elements are necessary or useful when generating narrative prose.

To study the effects of different textual effects upon the readers, we implemented five versions of the STORYBOOK story generator (Callaway & Lester 2001). Because a fully interleaved experiment would have required an excessive amount of time, we required each student to compare two versions of each story rather than all five versions. Each story was identical in plot, content, and form, but differed in terms of propositions per sentence, grammatical fluency, or choice of lexical forms. The results of the study show that the participants were highly discriminative of the texts which they read, preferring some versions over others. The readers most strongly dispreferred narratives lacking important grammatical structures and greatly dispreferred those with a small number of propositions per sentence. These results have important implications for the design of literacy software.

The STORYBOOK Narrative Prose Generator

STORYBOOK is a narrative prose generator that produces narratives in the Little Red Riding Hood domain. To write stories, STORYBOOK takes a narrative plan consisting of the actors, scenes, props and temporally ordered events and descriptions as input from a narrative planner. It then evolves that narrative plan into the final text seen by the reader using a sequence of architectural components:

- *Discourse History*: When populating a story with information from a conceptual network, noun phrases must be marked for indefiniteness if they have not yet been mentioned in the story or if they are not visually available references to the character or narrator in focus. Furthermore, frequently repeating noun phrases can be pronominalized to avoid sentences like “Grandmother knew that Grandmother had asked Grandmother’s daughter to send some cakes to Grandmother” rather than “Grandmother knew she had asked her daughter to send her some cakes.” A discourse history tracks noun phrase concepts and allows them to be marked for definiteness or pronominalization.

- *Sentence Planner*: A sentence planner maps characters, props, locations, actions and descriptions to concrete grammatical structures in a sentential specification. Thus in the example just mentioned, “grandmother” is mapped to the main subject while “know” is mapped to the main verb, etc.

- *Revision*: Because narrative planners create their content as a series of single proposition sentences, a revision component is usually introduced to *aggregate* those small sentences (protosentences) into larger multi-proposition sentences. It is usually assumed that these larger sentences will be more readable and less choppy or visually jarring. For example, “The wolf saw her” and “She was walking down the path” might be aggregated to produce “The wolf saw her walking down the path.”

- *Lexical Choice*: Narrative planners also tend to create sentences that frequently repeat the same lexical items due to efficiency concerns. To combat this, a lexical choice component performs local search to determine when one lexical item can be replaced by another. Thus instead of character dialogue where characters always introduce utterances with “said”, that lexical item can be replaced by “mentioned”, “whispered”, “replied”, etc.

- *Surface Realizer*: Once the lexical and structural content of a set of sentences has been determined, they must be converted to text. This is accomplished by checking to make sure that each sentence is grammatical, imposes linear constraints, and adds morphological changes as necessary. The result is text which can be sent to a word processor, a web browser, or saved as a text file.

The existence of these architectural modules allowed us to conduct an *architectural ablation* experiment. By selectively removing a component, the resulting text of a story will be changed in some way. The sentence planner and surface realizer are vital components; without them text cannot be produced at all. However, removing the other elements will result in text that we expect to be degraded in some fashion. Thus without the discourse history, the system will be unable to produce pronouns in a reliable way or appropriately mark nouns for definiteness. Without the revision component, the system will produce a minimal number of propositions per sentence due to the lack of clause aggregation. Finally, removing the lexical choice module will result in a decrease in the variability of the lexical forms of verbs or nouns.

Given these three architectural modules, there are 2^3 or 8 possible pairwise comparisons between the presence or absence of each component when used to produce a narrative:

1. All three components are used.
2. Only the revision module is unused.
3. Only the lexical choice module is unused.
4. Only the discourse history module is unused.
5. Only the revision module is used.
6. Only the lexical choice module is used.
7. Only the discourse history module is used.
8. None of the three components are used.

Due to the constraints on the logistics of the evaluation process, we decided to utilize only five of those pairwise comparisons: the two all-or-none approaches and the three approaches where one specific architectural module is ablated. The remaining three unused approaches would evaluate the enhancement that each module adds to the whole rather than what is missing when each is removed. We contend this approach leads to a slightly more effective comparison, because as more modules are removed from the generation process, the resulting prose becomes progressively less desirable and thus unwanted effects from the absence of multiple architectural modules might overlap and affect a test subject's experience in ways that could not be teased apart when analyzing the data.

The ablation of these architectural modules can have a significant impact in text quality, even over very small text segments, as is shown in the following excerpts:

- Complete (Version A), with revision, lexical choice, and discourse history all turned on:

  > She had not gone far when she met a wolf.
  >
  > "Hello," greeted the wolf, who was a cunning looking creature. He asked, "Where are you going?"
  >
  > "I am going to my grandmother's house," she replied.

- No Revision (Version B), with lexical choice and discourse history turned on:

  > She had not gone far. She met a wolf.
  >
  > "Hello," greeted the wolf. The wolf was a cunning looking creature. He asked, "Where are you going?"
  >
  > "I am going to my grandmother's house," she replied.

- No Lexical Choice (Version C), with revision and discourse history turned on:

  > She had not gone far when she met a wolf.
  >
  > "Hello," said the wolf, who was a cunning looking creature. He said, "Where are you going?"
  >
  > "I am going to my grandmother's house," she said.

- No Discourse History (Version D), with revision and lexical choice turned on:

  > Little Red Riding Hood had not gone far when Little Red Riding Hood met the wolf.
  >
  > "Hello," greeted the wolf, who was the cunning looking creature. The wolf asked, "Where is Little Red Riding Hood going?"
  >
  > "Little Red Riding Hood is going to Little Red Riding Hood's grandmother's house," replied Little Red Riding Hood.

- Empty (Version E), with revision, lexical choice, and discourse history all turned off:

  > Little Red Riding Hood had not gone far. Little Red Riding Hood met the wolf.
  >
  > "Hello," said the wolf. The wolf was the cunning looking creature. The wolf said, "Where is Little Red Riding Hood going?"
  >
  > "Little Red Riding Hood is going to Little Red Riding Hood's grandmother's house," said Little Red Riding Hood.

1. On an absolute scale of how good fairy tales should be in general, evaluate the story on an A–F scale (A, B, C, D, F).
2. Style: Did the author use a writing style appropriate for fairy tales?
3. Grammaticality: How would you grade the syntactic quality of the story?
4. Flow: How well did the sentences flow from one to the next?
5. Diction: How interesting or appropriate were the author's word choices?
6. Readability: How hard was it to read the prose?
7. Logicality: Did the story omit crucial information or seem out of order?
8. Detail: Did the story have the right amount of detail, or too much or too little?
9. Believability: Did the story's characters behave as you would expect?

Figure 1: Grading factors presented to readers

Evaluation Methodology

To test the StoryBook system, we created a modestly sized narrative planner (implemented as a finite state automaton containing approximately 200 states), enough to produce two stories comprising two and three pages respectively. Furthermore, we fixed the content of those stories and ran five different versions of StoryBook on each one: (A) all three components working, (B) revision turned off, (C) lexical choice turned off, (D) the discourse history turned off, and finally (E) a version with all three components turned off. This resulted in ten total narratives which we presented to our test subjects using the grading factors found in Figure 1. While versions were different in the sense that certain modules were either ablated or not, the two stories differ because they were created from two different finite state automata. Thus story #1 potentially has different characters, different events and properties, and different props than story #2 has.

A total of twenty students were selected from North Carolina State University's Departments of English and

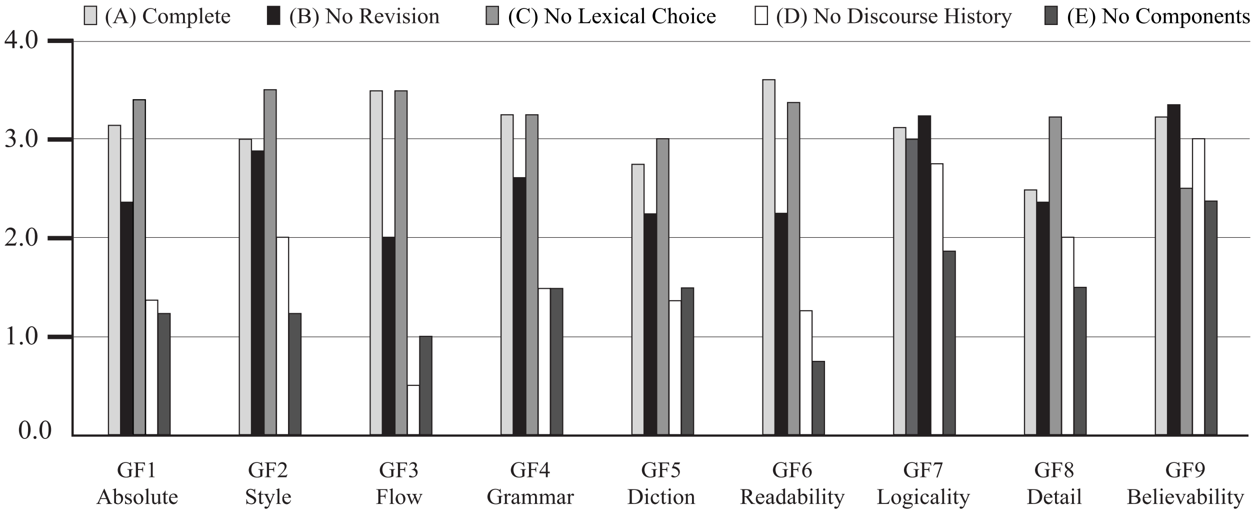

Figure 2: Means for Story #2: 4.0 scale, 8 evaluations per Version × Grading Factor × Story

Communication via first-come first-serve email notices. All of the students were registered in upper division or graduate courses in those departments. Each subject was asked to read the directions and ask for clarifications before the evaluation proceeded and was randomly assigned their evaluation task. Subjects were not informed prior to their completion of the questionnaire that the narratives were produced by computer program. Subjects were paid $25.00 for their participation.

Because each subject compared two versions of story #1 to each other and two versions of story #2 to each other, every subject saw a total of four narratives. To prevent subjects from evaluating the same types of stories in succession, we devised the following policy:

1. Each subject read four distinct story versions out of the total of five, two from each story (*e.g.*, subject #1 read versions A and B from story #1, and versions D and E from story #2). No subject read the same version twice.

2. Each version was read by the same total number of subjects (*i.e.*, each version of every story was read by 8 separate subjects).

3. Each pairwise comparison of different versions was read by two separate subjects (*e.g.*, subjects #1 and #11 both read versions A and B of story #1 and versions D and E of story #2).

4. For each pair of students reading the same two versions, the narratives were presented in opposite order (*e.g.*, subject #1 read version A first and then version B, while subject #11 read version B first followed by version A).

5. Students were randomly assigned narrative versions on a first-come first-serve basis; all students performed their evaluations within 3 hours of each other at a single location.

Subjects graded each narrative following the instructions according to an A–F scale, which we then converted to a quantified scale where A = 4.0, B = 3.0, C = 2.0, D = 1.0, and F = 0.0. The resulting scores were then tallied and averaged. The means for both stories are shown in Figure 2.

To determine the quantitative significance of the results, we performed an ANOVA test over both stories. The analysis was conducted for three independent variables (test subject, story, and version) and nine grading factors (labelled GF1 – GF9, as described in Figure 1). Because not all possible grading combinations were performed (only 80 observations, or 20 x 2 x 2, out of a possible 200, or 20 x 2 x 5, due to crossover and time constraints), we performed the mixed procedure analysis. Interactions between variables were only significant for grading factor #9 at 0.0300 for story∗version.

The results of the ANOVA analysis point to three significant classes of narratives due to the architectural design of the narrative prose generator. Table 1 indicates that the most preferred narrative class, consisting of versions A & C, were not significantly different from each other overall while they did differ significantly from all other versions (although there were similarities in particular grading factors such as GF2, *style*, between versions A & B). Interestingly, the affinity for versions A & C is strongly correlated for story #2 (Figure 2) but only weakly for story #1. A two-tailed paired t-test evaluating this difference illustrated that versions A & B were not significantly different when only story #1 was considered, but were significantly different in story #2. The opposite was true for versions A & C when the scores for each story were compared individually, even though the combined scores indicated versions A & C were not significantly different overall.

Grading Factors	GF1	GF2	GF3	GF4	GF5	GF6	GF7	GF8	GF9	ALL
COMPLETE VS. NO REV.	n.s.	n.s.	$**$	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
COMPLETE VS. NO L. C.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
COMPLETE VS. NO D. H.	$**$	$*$	$**$	$**$	$**$	$**$	n.s.	$*$	n.s.	$**$
COMPLETE VS. NOTHING	$**$	$*$	$**$	$**$	$**$	$**$	n.s.	n.s.	$*$	$**$
NO REV. VS. NO L. C.	$*$	n.s.	$**$	$*$	$*$	$*$	n.s.	n.s.	n.s.	$**$
NO REV. VS. NO D. H.	$**$	$*$	$**$	$**$	$*$	$**$	n.s.	n.s.	n.s.	$**$
NO REV. VS. NOTHING	$**$	n.s.	$*$	$**$	n.s.	$**$	n.s.	n.s.	$*$	$**$
NO L. C. VS. NO D. H.	$**$	$**$	$**$	$**$	$**$	$**$	$*$	$**$	$*$	$**$
NO L. C. VS. NOTHING	$**$	$**$	$**$	$**$	$**$	$**$	$*$	$**$	$**$	$**$
NO D. H. VS. NOTHING	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Table 1: Combined significance values (with Bonferroni adjustment): $*$ = $p < 0.01$, $**$ = $p < 0.001$

Discussion

Indisputably, versions D & E form the least preferred narrative class, differing quite significantly from all other versions while not differing significantly from each other. Because the architectural commonality between these two versions was the lack of a discourse history (corresponding to a lack of grammatical conformity to the expected norm, especially lack of appropriate pronominalization) while versions A, B, and C all utilized a discourse history, we conclude that this architectural component is extremely important in the design of a narrative prose generator and that any symbolic pipelined narrative prose generation system will suffer tremendous degradation in prose quality if a discourse history component is not present. In addition, we conclude that in future ablation experiments, if there is no other methodology for introducing pronominalizations, it is not even desirable to include the discourse history module as one of the components available for ablation. Effects of pronominalization and topicalization were previously studied by Hoover (1997) although that work focused on recall rates while we concentrate on expressed preferences.

As predicted in advance, the full version (Version A) scored quite well while versions lacking a discourse history (Versions D & E) scored quite poorly. A surprise in the results of the analysis was the mild preference subjects had for the version missing the lexical choice component (Version C) over the full-fledged version. While related work on word choice in spontaneous dialogues has concluded that dialogue participants tend to converge onto a limited set of words (Brennan 1996), fictional narrative by and large does not reflect the spontaneity and task-orientation reflected in such dialogues.

Upon analysis of the comments in the evaluations specifically comparing versions A & C, it became clear that one principal reason was the test subjects' belief that the increased lexical variation might prove too difficult for children to read (even though we provided no indication that the target audience was children) and thus Version A compared less favorably to Version C due to the more complex and varied words it contained. It is not clear whether a lexical choice component would play a much more significant role in subject matter where the audience was more mature.

The fact that Version B scored less favorably compared to Versions A and C indicates that revision is an important aspect of narrative prose generation. Test subjects frequently commented that Version B was "very choppy" or "didn't seem to have good grammar". These comments can be accounted for by the two main functions of the revision component: joining small sentences together and combining sentences with repetitive phrases together while deleting the repetitions. This is related to previous work in reading comprehension on propositional content. Such research (Kintsch & Keenan, 1973) has shown that reading rate increases as the number of propositions per sentence increases. Here, however, we have shown that a larger number of propositions per sentence is preferred more than a small number of propositions per sentence, although there would certainly be an upper limit.

Another important note is that there is a difference among the grading factors themselves. Grading factors 2-7 (style, flow, grammar, diction, readability and logicality) directly relate to elements governed by the parameters and rules of the various architectural components of the narrative prose generator. However, grading factors #8 and #9 (detail and believability) are more closely related to the content of the plot line, and as such could be expected to remain relatively constant since the content of the narratives was held constant across all versions of each story. Given that the perceptions of the test subjects might have "carried over" from their responses to previous questions, a future evaluation might randomize the order in which these questions are asked to see if this effect persists.

Finally, there appears to be a link between the appeal of the story content itself and the increase in the absolute (GF #1) and total means for versions A, B, and C. Story #1 is a "classic" Brothers' Grimm fairy tale in the sense that it typically has a gruesome ending that serves as a behavioral warning to young children. Thus our story #1 ends with the wolf devouring Little Red Riding Hood and her grandmother. More modern stories have happier endings, however, and this is reflected in our story #2 which ends with a woodcutter killing the wolf and extracting the unharmed Little Red Riding Hood and her grandmother from the wolf's stomach. A large number of our test subjects, worried about the potential impact on children, complained about the "horrible" ending of story #1 in their written comments and this reader bias appears to have affected the overall grading scores.

Future Work

The existence of a computational system for generating complete narratives while providing access to the fundamental linguistic structure offers superb opportunities for future experimentation. Very fine-grained manipulation of texts becomes possible on a large scale; for example, within the discourse history, it is possible to run ablation experiments involving subject pronouns vs. object pronouns, correct vs. incorrect reflexive pronouns, random vs. ambient definite noun phrase marking, among many others.

Acknowledgements

The authors wish to thank Joy Smith of NC State University for her help with the statistical analysis and the anonymous reviewers for their extremely helpful comments. Support for this work was provided by the IntelliMedia Initiative of North Carolina State University.

References

Allen, J., Miller, B., Ringger, E., & Sikorski, T. (1996). Robust understanding in a dialogue system. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, (pp. 62–70). Santa Cruz, CA.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*. Philadelphia, PA.

Callaway, C., & Lester, J. (1997). Dynamically improving explanations: A revision-based approach to explanation generation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, (pp. 952–958). Nagoya, Japan.

Callaway, C., & Lester, J. (2001). Narrative prose generation. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, in press. Seattle, WA.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, **48**: 163–189.

Hoover, M. L. (1997). Effects of textual and cohesive structure on discourse processing. *Discourse Processes*, **23**: 193–220.

Kintsch, W. & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**:257–274.

Lang, R. R. (1997). *A formal model for simple narratives*. Doctoral Dissertation, Department of Computer Science, Tulane University. New Orleans, LA.

Lebowitz, M. (1985). Story-telling as planning and learning. *Poetics*, **14**(3): 483–502.

Lester, J. & Porter, B. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, **23**(1): 65–101.

Meehan, J. (1977). Tale-Spin, an interactive program that writes stories. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Cambridge, MA.

Robin, J. & McKeown, K. (1995). Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, **85**(1–2).

Smith, R. & Hipp, D. R. (1994). *Spoken natural language dialog systems*. Cambridge, Massachusetts: Oxford University Press.

Young, R. M. (1999). Using Grice's Maxim of Quantity to select the content of plan descriptions. *Artificial Intelligence*, **115**: 215–256.