Effective Scenario Designs for Free-Text Interactive Fiction

Margaret Cychosz[1], Andrew S. Gordon[2(✉)], Obiageli Odimegwu[2], Olivia Connolly[2], Jenna Bellassai[3], and Melissa Roemmele[2]

[1] University of California, Berkeley, CA, USA
mcychosz@berkeley.edu
[2] University of Southern California, Los Angeles, CA, USA
{gordon,roemmele}@ict.usc.edu, {odimegwu,oconnoll}@usc.edu
[3] Oberlin College, Oberlin, OH, USA
jbellass@oberlin.edu

Abstract. Free-text interactive fiction allows players to narrate the actions of protagonists via natural language input, which are automatically directed to appropriate storyline outcomes using natural language processing techniques. We describe an authoring platform called the Data-driven Interactive Narrative Engine (DINE), which supports free-text interactive fiction by connecting player input to authored outcomes using unsupervised text classification techniques based on text corpus statistics. We hypothesize that the coherence of the interaction, as judged by the players of a DINE scenario, is dependent on specific design choices made by the author. We describe three empirical experiments with crowdsourced subjects to investigate how authoring choices impacted the coherence of the interaction, finding that scenario design and writing style can predict significant differences.

1 Free-Text Interactive Fiction

Among the various benchmarking evaluations used in Artificial Intelligence research, the Choice Of Plausible Alternatives (COPA) evaluation [9] is particularly interesting in its relationship to interactive digital storytelling. In each item of this 1000-question evaluation, software systems are presented with an English-language *premise* and two *alternatives*, and asked to select which of the two is more plausibly the causal consequence of the premise (or in some questions, its antecedent). In all questions, both alternatives are possibly the next event in the narrative context of the premise, but only one is unanimously judged as more plausible by multiple human raters. An example question from the COPA development set is as follows:

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

N. Nunes et al. (Eds.): ICIDS 2017, LNCS 10690, pp. 12–23, 2017.
https://doi.org/10.1007/978-3-319-71027-3_2

Despite its simplicity, success on the COPA evaluation has been a challenge for researchers in both commonsense reasoning and natural language processing. For commonsense reasoning researchers, the broad-domain of the situations presented in the questions has proved to be difficult to tackle with formal modeling methods, as has the use of natural language representations. For natural language processing researchers, the problem is that COPA includes no training data for use with familiar supervised machine learning approaches. Indeed, the systems that have achieved success on the COPA evaluation have all employed unsupervised approaches that gather statistical information about how words co-occur in very large, broad-domain text corpora [1,3,4,9].

The struggles of researchers working on the COPA evaluation have much in common with those seen in interactive digital storytelling. Here, researchers continually strive to increase the agency and free-will of players, and provide authors with the technical tools to create interactive experiences across an increasingly broad range of narrative domains. As with COPA questions, the central concern is algorithmically determining *what happens next* as a consequence of a player's action. In addition, the goal to support rich agency in increasing open domains will be difficult to meet using formal models of fictional story worlds and their causal mechanisms. Noting the parallels between COPA and interactive digital storytelling, we began to explore whether approaches that performed well on the COPA evaluation could be used *directly* as an engine for interactive fiction.

To test this idea, we built the Data-driven Interactive Narrative Engine (DINE), a web-based platform for authoring and deploying textual interactive fiction. DINE scenarios are structured as an interconnected set of *pages*, each of which is a variant of a COPA question consisting of a *setup* and an arbitrary number of alternatives *(outcomes)*, which are automatically selected and shown as the causal consequences of the player's own free-text narration of their actions (the *premise*). For example, the previous COPA question might be transformed into a DINE page in the following manner:

(setup)	On his porch, I heard the football game on my neighbor's television.
(input)	`>`
(outcome)	My neighbor invited me in. We sat down to watch the game [...]
(outcome)	Through the window I could see him on his couch. [...]
(outcome)	With my ear to the door, I heard him cheering for his team. [...]
(outcome)	Returning home, I found myself locked out of my house. [...]

In this example DINE page, the player reads the page setup and types what they would do in the situation, e.g., `I knocked on my neighbor's door`. The system then processes the input as the premise of a COPA question, selecting and displaying the outcome that is most plausibly the causal consequence of the player's action. Authors can link these outcomes to subsequent DINE pages or to endings, allowing for arbitrarily complex branching storylines [7]. Alternatively, an outcome can be displayed without advancing the storyline, prompting the player to narrate another action in the same context. In these cases, DINE will select the next most plausible outcome that has not already been shown to

the player. When the player feels that the shown outcome is incoherent, DINE provides a *huh?* button that replaces the outcome with the next one in the ranked list.

To automatically rank the authored outcomes, we used a top-performing COPA system that computes the average Pointwise Mutual Information (PMI) between words in the player's input and the first six words of each outcome, where the pairwise PMI statistics are computed by processing a corpus of millions of nonfiction personal stories from Internet weblogs [3]. Although other systems have demonstrated improvements on the COPA evaluation that are statistically significant [4], the differences are marginal, and not likely large enough to be apparent to players of a DINE interactive scenario. Instead, we believe that more substantive differences in the coherence of player interactions are going to stem from how the scenarios are written. Well-crafted DINE scenarios will have compelling setups that establish expectations about the space of possible outcomes, and compelling outcomes that adequately cover this space while advancing rich storylines. Coherence may furthermore be affected by how outcomes relate to each other on a page, or even by stylistic choices of the author in presenting the voice of the narrator and storyline characters. Rather than focus on technology, this paper explores how choices made by the authors of DINE scenarios can promote coherent interactive storytelling experiences.

We hypothesize that the coherence of the interaction, as judged by the players of a DINE scenario, is dependent on specific design choices that the author makes. In this paper, we describe our efforts to empirically investigate this hypothesis through a series of three large-scale experiments with crowdsourced subjects.

2 Scenario Design Patterns

We employed two full-time writers for ten weeks, tasked with authoring a collection of DINE scenarios for use in our experiments. During this period, these writers authored 100 new DINE scenarios, consisting of a total of 210 pages. Roughly half of these scenarios were single-page interactions, with the remainder consisting of four or fewer pages. The topics and genres of these fictional scenarios were left up to the authors, and included science-fiction adventures, psychological thrillers, romantic comedies, and skills-training situations. We encouraged the authors to explore the space of possible interaction styles afforded by the DINE approach, and conducted daily discussions of the merits of design decisions.

Several surprising design patterns emerged from the authors' creative efforts. We organized the design space along two dimensions, namely how the page outcomes were related to each other and how the page setup prompted players to type certain actions. With respect to the page outcomes, our writers made them either *order-dependent* or *order-independent*. Our initial intuition was that DINE writers should prefer order-independent outcomes, where the coherence of each outcome's text would be high regardless of which of the other outcomes had already been shown to the player. In this model, players might try a number of

actions to move the story forward before transitioning to a new page. Crucially, none of these actions would change the current context enough to invalidate the coherence of other outcomes. An example of this design was seen in a page where the player is trying to stay awake while driving a cargo truck on the highway at night, as follows:

(input) `> I rolled down the window`

(outcome) As I rolled down the windows to usher in the night air, I felt disheartened. It was muggy out. Lukewarm and goopy like soup left untended and allowed to congeal. It settled around me, heavy, making me even more drowsy than before. The road swam, slow and sluggish, before my eyes.

(input) `> I drank my energy drink`

(outcome) As I gulped the energy drink down in long, greedy pulls, I felt hopeless. I had tried this before. Coffee. Five-hour Energy. Even caffeine pills once. It would work but only for an hour or two. [...]

(input) `> I finished my sandwich`

(outcome) Eating the sandwich only made me more drowsy. I felt it sitting heavy at the bottom of my belly. [...]

(input) `> I turned on the radio`

(outcome) The radio blared country music into the silence of the car. [...] It was all just a dull ringing in the air around me. My eyes kept easing shut and before snapping open again.

The interactive experience was much different on pages with order-dependent outcomes, i.e. where outcomes would only be coherent to the player if they had already seen one or more of the other outcomes on the page. This design pattern was frequently used by our writers to lead players through a specific sequence of behaviors, with each new outcome suggesting the behavior that the author expected next from the player. In this model, the text would often only make sense if the player was a willing participant — any creative deviations on their part would derail the writer's narrative. An example of this design was seen in a page where the player is exploring an empty house while dog-sitting for a neighbor, as follows:

(setup) The dogs immediately rushed upstairs. I didn't want them to get into trouble or break anything else. I followed them up to the third floor. On the ground of the hallway was a shattered picture frame.

(input) `> I looked at the picture`

(outcome) Picking the frame up, I looked at the faded black and white photo of a family in front of the farmhouse. Mr. Larson had shown me the photo before. Henry sat on the porch with his two brothers and two sisters while their parents stood behind them. Sammy the dog began licking water off the hardwood floor.

(input) `> I wondered why the floor was wet`

(outcome) The hardwood floor hallway was wet, as if someone dripping in water had walked along the hall. The water stopped at the fourth room down the hall.

(input) `> I went down the hall`

The strong order-dependence of the outcomes produces an interactive narrative that is essentially linear — the player's textual input only serves to advance the storyline to the next passage of written text. This model questions the role of (perceived) free-will in interactive storytelling, emphasizing instead a cooperation between author and player. The player's job is to figure out what text the author expects the player to type, and the writer's responsibility is to make it possible to do so given the storyline.

The second dimension in the space of page design patterns was how the page setup was written, specifically how it prompted the player to take certain actions. We observed three main types of setups in the pages written by our authors:

1. A **mystery** setup presents the player with some problem or puzzle that needs to be solved. The author expects the player to type actions that they believe will solve the mystery. Examples: *How do I get out of this locked room? How do I keep my teenage friends from starting a forest fire?*
2. A **decision** setup presents the player with a forced choice. The author expects the player to either choose one of the options, or gather more information to make the decision. Examples: *Should I tell my friend that she looks ridiculous in her new dress? Should I intervene when I see a parent harming a child?*
3. A **task** setup informs the player of exactly what they are expected to do. Examples: *I am in front of the queen and I am expected to bow. I just woke up and I am expected to do my morning stretches.*

The actual length of the setups written by our authors varied widely, from a single paragraph to the length of a chapter in a novel. Often the very long setups would include backstory and character development that was largely unrelated to the player's interaction. For example, the player may read the long, depressing tale of the protagonist's failed attempt to become a professional concert pianist, as setup for an interaction about a mysterious noise in the middle of the night. These DINE scenarios blurred the genre boundary between interactive fiction and traditional short stories, with the interactive component serving as an intermission in a linear text, rather than as a central focus of the work.

3 Authoring Experiments

Observing a broad range of scenario design patterns from the two writers, we sought to better understand the impact of these design decisions through a series of human-subject experiments. As a web-deployed application, the DINE platform affords the easy collection of player interaction data from crowdsourced workers on the Internet. We recruited crowdsourced workers to participate in

three controlled experiments, each targeting a different set of scenario characteristics. In the first experiment, we investigated the effect of the type of setup (*mystery*, *decision*, or *task*), using 25 scenarios across these three categories as stimuli. In the second experiment, we explored the role of setup length and outcome structure (*order-dependent* or *order-independent*) with stimuli that manipulated these characteristics in multiple versions of the same scenario. In the third experiment, we manipulated the tense (*past* or *present*) and dialogue style (*narrated* or *quoted*), and further investigated whether these factors affected the player's own writing style, and whether this in turn affected the accuracy of the underlying model for selecting outcomes.

3.1 Experiment 1: Setup Type

We observed three types of setups in the scenarios authored by the two writers (*mystery*, *decision*, or *task*). Each type creates different expectations about what sort of text is likely to be entered by the player, and may respond differently to unanticipated or creative player input. Mystery-type setups place the fewest expectations on the player's actions, and require the author to anticipate a broad variety of potential actions when crafting the scenario. In decision-type setups, the player is expected to take actions that correspond to implicit or explicit options, and the author must at least provide appropriate outcomes for the range of choices. In task-type setups, the player is expected to do one thing only, and the author focuses on appropriately responding to this action. We hypothesized that the coherence of the player experience, indicated by the dependent variables of huh-rate and coherence ratings, is determined by the degree to which the author was required to anticipate player creativity, i.e. that mystery-type setups would yield the most coherent scenarios, and task-type setups the least.

For this first experiment, $N = 393$ participants were recruited from an online crowdsourcing service (http://www.crowdflower.com). All participants were self-reported American English speakers living in the United States at the time of the experiment. Each participant completed one interactive DINE scenario and was compensated $1.00 USD. Total participation time was no more than 8 min. A total of 2368 user inputs were collected. No demographic information was collected from participants in this first experiment. Participants were redirected from the crowdsourcing website to the online website that hosts DINE scenarios. Participants were told that they would be interacting with a computer to tell a story and were also told how to generate alternative responses via the *huh?* button when presented with incoherent outcomes. The interactive scenario ended when the player had reached a terminal outcome, or when all available outcomes had been presented. Data from all unfinished scenarios were discarded from analysis. At the end of the experiment, participants completed a post-questionnaire where they rated the coherence of the interaction on a five-point Likert scale, answering the question "How coherent was your story?"

As experimental stimuli, we selected 25 DINE scenarios from the pool of 100 scenarios authored by the two writers, distributed across three categories of setup-type. Mystery-type setups, where the player is presented with some

Table 1. Summary statistics for Experiment 1

Setup type	Coherence rating		Huh-rate
	Median	Mean (SD)	Mean (SD)
Mystery	4	3.66 (1.17)	0.19 (0.07)
Decision	4	3.59 (1.26)	0.20 (0.06)
Task	3	3.18 (1.20)	0.25 (0.11)

problem or puzzle to solve, were the most prevalent in these scenarios (N = 14). Decision-type setups (N = 5) presented players with a forced choice, and task-type setups (N = 6) told the player exactly what they were supposed to do.

Table 1 summarizes the differences observed in coherence ratings and huh-rate across the three setup types. As per our hypothesis, mystery-type setups produced the highest coherence ratings and lowest huh-rate, whereas task-type setups produced the lowest coherence ratings and highest huh-rate. We found that setup-type (mystery, decision, task) was a significant predictor for coherence ratings ($\beta = -0.61811$, $p = 0.038$), fitting a cumulative link mixed model (ideal for ordinal dependent variables). However, setup-type was not a significant predictor of huh-rate, analyzed using a generalized linear mixed effects model. Furthermore, we observed that our two dependent measures were not directly correlated in these 25 scenarios (cor = 0.05). To better control the individual factors that may determine these two measures in our subsequent experiments, we changed our approach to use multiple variations of a single DINE scenario as our experimental stimuli.

3.2 Experiment 2: Setup Length and Order Dependence in Outcomes

In Experiment 2 we examined the role of two additional structural characteristics that varied in the writers' scenarios, namely length of the setup and order dependencies in the outcomes. Our hypothesis was that longer setups would lead to more coherent interactions, as more text would afford more opportunities for the author to establish the goals and disposition of the protagonist; knowing who they were should help players know what they should do. We also hypothesized that DINE scenarios with order-dependent outcomes would be less robust to player creativity and more susceptible to failures due to classification errors, resulting in lower coherence ratings and higher huh-rates. Our approach in Experiment 2 was to experimentally manipulate these factors as independent variables, namely by having our writers craft four versions of a single scenario for use as stimuli in a 2×2 experimental design.

For the second experiment, N = 200 additional English-speaking America-residing participants were recruited from the same online crowdsourcing service. Mean participant age was 32.08 (SD = 10.66). Each participant completed one interactive DINE scenario and was compensated $1.00 USD. Total participation

time was no more than 8 min. A total of 711 user inputs were collected. Participants were eliminated from the analysis for the following reasons: performed the experiment twice (second performance deleted) or completed demographic information but did not perform experiment. This left the analysis with $N = 190$ participants. Only $N = 135$ participants provided the post-hoc Likert-scale coherence rating. Participants who did not complete the post-questionnaire were not eliminated from the analysis in Experiment 2. Participants interacted with the DINE system as instructed in Experiment 1.

For stimuli, we selected a single DINE scenario, an absurdist psychological-horror story about a home invasion, as the basis for creating four experimental variations. Each of these four variations was created by the original writer of the scenario, with instructions about two dimensions of variation. First, we varied the length of the setup. Two variants were given very long setups of approximately 1,000 words, providing the player with a rich backstory about the scenario protagonist that details the difficult life events that preceded the night when a mysterious noise in her bedroom wakes her up. The other two variants provide only a brief setup of approximately 50 words, describing awakening to the sound of a mysterious noise. Likewise, we varied the order-dependence of the scenario outcomes. Two variants were written with high order-dependence, where each outcome was suggestive of the next action expected of the player. The other two variants were written with order-independent outcomes, where the outcomes would be coherent regardless of the order they were read by the player.

Table 2 summarizes the differences observed in coherence ratings and huh-rate across each of the two sets of variables. The results show higher coherence ratings for scenarios with a long setup and order-independent outcomes, and lower huh-rates for scenarios with a short setup and order-independent outcomes. The order dependence of outcomes was a significant predictor of coherence ratings ($\beta = -0.9530$, $p = 0.020$), fitting a cumulative link mixed model, but setup length was not a significant predictor. Neither setup length nor outcome order-dependence had a significant effect on huh-rate, as determined by a two-way ANOVA. Our analyses of huh-rate statistics suggests that it may be too coarse of a dependent measure for evaluating the impact of author design choices. We addressed this issue in Experiment 3 by conducting a more thorough analysis of the coherence of individual interactions.

Table 2. Summary statistics for Experiment 2

Variable	Coherence rating		Huh-rate
	Median	Mean (SD)	Mean (SD)
Long setup	3	3.43 (1.10)	0.25 (0.35)
Short setup	3	3.32 (1.23)	0.21 (0.32)
Order-dependent	3	3.14 (1.17)	0.20 (0.32)
Order-independent	4	3.55 (1.24)	0.26 (0.35)

3.3 Experiment 3: Setup Tense and Dialogue Style

In Experiment 3 we investigated how the author's writing style affected the coherence of DINE scenarios, specifically looking at the variables of tense (present or past) and the way that dialogue is written (quoted or narrated) in the setup and outcomes of a scenario. In traditional interactive fiction it is common to author scenario text in the present tense second-person voice, e.g. *You are in a maze of twisty little passages, all alike* [6]. In contrast, we observed that our two writers wrote DINE scenarios exclusively in the first-person past tense. Likewise, our writers freely mixed the use of quoted dialogue (direct speech) and narrated dialogue (indirect speech) when describing conversations between storyline characters. We hypothesized that these stylistic variations would be factors in the coherence of DINE scenarios, reasoning that players would adapt their own writing style to match that of the authors, and that these differences might affect the performance of the unsupervised text classifier that underlies the DINE software.

We investigated tense and dialogue style as independent variables in a 2×2 experimental design, using four variations of a single DINE scenario as our experimental stimuli. An additional $N = 200$ English-speaking America-residing participants were recruited from the same online crowdsourcing service. Mean participant age was 32.1 ($SD = 11.50$). Each participant completed one interactive DINE scenario and was compensated $1.00 USD. Total participation time was no more than 8 min. A total of 2536 user inputs were collected. All participants were required to complete the post-questionnaire in Experiment 3. Participants interacted with the DINE system as instructed in Experiments 1 and 2.

As stimuli, we selected a single DINE scenario as the basis for creating four experimental variations. In this story, the player is a school teacher struggling to deal with a male student who is both abusive and abused, written as a mystery-type setup with order-independent outcomes. Each of these four variations was crafted by the two writers, with instructions about two dimensions of variation. First, we varied the tense, writing two variations entirely in the first-person past tense (*"I stood nervously outside the principal's office*) and two in the first-person present tense (*I stand nervously outside of the principal's office*). Second, we varied the style of dialogue, with two variants with quoted dialogue (*"We do need to do something."*) and two with narrated dialogue (*He agrees with me that we do need to do something*).

Table 3 summarizes the differences observed in coherence ratings and huh-rate across each of the two sets of variables. The results show higher coherence ratings for past tense scenarios and narrated dialogue, and lower huh-rates for present tense scenarios and narrated dialogue. Fitting a cumulative link model we found that neither variable was a significant predictor of coherence ratings ($p > 0.05$), but observed a tendency for the variable of dialogue ($\beta = -0.4455$, $p = 0.0781$). Specifically, a negative beta coefficient for quoted dialogue indicates that as quoted dialogue is used, the coherency ratings decrease. Analysis of variance (two-way ANOVA) showed a significant effect for the dialogue variable on huh-rate ($F(1,211) = 11.1375$, $p < 0.001$) but not for tense.

Table 3. Summary statistics for Experiment 3

Variable	Coherence rating		Huh-rate
	Median	Mean (SD)	Mean (SD)
Past tense	3	3.36 (1.15)	0.27 (0.20)
Present tense	3	3.22 (1.29)	0.25 (0.26)
Quoted dialogue	3	3.16 (1.21)	0.32 (0.29)
Narrated dialogue	4	3.45 (1.20)	0.20 (0.23)

Effect of Writing Style on Player's Language Use. Given the significant effect of dialogue type on huh-rate and a tendency in that direction for coherence rating, we conducted two additional analyses to determine (1) if the type of dialogue the authors used affected user dialogue choice and (2) if story tense affected the tense that users chose. To conduct this analysis, user inputs from Experiment 3 were hand-annotated by a single annotator with formal training as a linguist. This set was first filtered to remove user inputs that were unintelligible or not narrative text ($N = 200$), as well as inputs for which there was no possible coherent outcome in the stimuli scenarios ($N = 588$). The remaining inputs were annotated for tense (present or past) and dialogue style (narrated or quoted). Some player inputs had ambiguous tense ($N = 1{,}087$) or did not contain dialogue ($N = 246$) and were discarded from analysis. This resulted in $N = 1{,}500$ examples of user input for dialogue analysis and $N = 661$ for tense analysis.

Analyzing this data by fitting binary logistic regression models, we found that the writer's choice of tense was a strong predictor of the user's choice of tense ($\beta = 0.5590$, $p < 0.0001$), and that the writer's choice of dialogue style was a strong predictor of the user's choice of dialogue style ($\beta = 2.1508$, $p < 0.0001$).

Effect of Player's Language Use on Classification Accuracy. Given the effect of scenario tense and dialogue style on the player's language use, we next investigated whether these language characteristics affected the performance of the underlying model that is used to select outcomes, described in Sect. 1. To conduct this analysis, Experiment 3 user inputs were hand-annotated with the scenario outcome that constituted the most coherent response, as judged by a single annotator trained as a linguist, and compared with the outcome actually selected by the underlying DINE text classifier. All user inputs from the previous analysis were annotated, with the addition of those interactions labeled as having no dialogue or ambiguous tense (these were eliminated in the previous analysis). This resulted in a total of $N = 1{,}742$ user inputs, each assigned to one of three dialogue style classes (no dialogue, narrated dialogue, or quoted dialogue), one of three tense classes (ambiguous tense, present tense, or past tense), and the correctness of its classification.

A binary logistic regression was fit to predict the classification (correct or incorrect) with the predictors of tense and dialogue style. Dialogue style was a significant predictor ($\beta = 0.6240$, $p = 0.0217$), where an incorrect classification is

more likely for user input of the no dialogue class. Tense showed an insignificant effect ($\beta = 0.5863$, $p = 0.0847$) with a tendency for present-tense input to predict incorrect classifications.

In summary, Experiment 3 identified a significant effect of the author's dialogue style on huh-rate and a tendency in that direction for coherence rating. Together, the two subsequent analyses hint at the causal mechanisms involved. We see that players are likely to match the author's dialogue style and tense when typing their intentions into DINE scenarios, and that players' choice of dialogue style has a significant effect on the ability of the underlying model to select the most appropriate outcome (bias against no dialogue), and a tendency for tense, as well (bias against present tense).

4 Discussion

From the player's perspective, interactions with DINE are not markedly different from previous free-text interactive digital storytelling prototypes. From the author's perspective, however, DINE greatly reduces the amount of development effort required to successfully process natural language player input. Previously, language processing pipelines have required knowledge-based parsers backed by rich domain models [5,8], or the collection and annotation of copious amounts of player input for use as training data [2,12]. In many ways, DINE resembles recent attempts at case-based interactive digital storytelling [10,11], where large corpora of narrative texts are used to make predictions about what happens as a result of the player's actions. DINE differs from these systems in the way that text corpora are exploited; instead of assembling new stories from multitudes of contributing authors, DINE uses corpus statistics to select its contributions from those written by a single author. DINE removes the technical aspects of language processing from the authoring process, shifting the focus toward the more familiar task of telling good stories.

From a research perspective, the ability to rapidly author new scenarios affords new opportunities for empirical evaluations, where variations in the scenario are the experimental manipulation. Experimental manipulations with large subject pools are uncommon in interactive storytelling research precisely because of high scenario development costs, in both time and expertise. By reducing these costs, we made a number of new findings concerning free-text interactive fiction. We found setup type, outcome order-dependence, and (possibly) dialogue style were predictors of the coherence of player's interactions. We found that players match the writing style of authors with respect to tense and dialogue style, and that these changes were predictive of the performance of the underlying model for selecting outcomes. These findings provide guidance to future authors of DINE scenarios, and encourage future exploration of novel designs and algorithms that further support free-text interaction in interactive digital storytelling.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 1560426. The projects or efforts depicted were or

are sponsored by the U.S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Goodwin, T., Rink, B., Roberts, K., Harabagiu, S.: UTDHLT: COPACETIC system for choosing plausible alternatives. In: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Canada (2012)
2. Gordon, A., van Lent, M., van Velsen, M., Carpenter, P., Jhala, A.: Branching storylines in virtual reality environments for leadership development. In: Proceedings of the Sixteenth Innovative Applications of Artificial Intelligence Conference (IAAI-2004), San Jose, CA (2004)
3. Gordon, A.S., Bejan, C., Sagae, K.: Commonsense causal reasoning using millions of personal stories. In: Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-2011), San Francisco, CA (2011)
4. Luo, Z., Sha, Y., Zhu, K.Q., Hwang, S.W., Wang, Z.: Commonsense causal reasoning between short texts. In: 15th International Conference on Principles of Knowledge Representation and Reasoning (KR-2016), Cape Town, South Africa (2016)
5. Mateas, M., Stern, A.: Integrating plot, character and natural language processing in the interactive drama facade. In: Proceedings of Technologies for Interactive Digital Storytelling and Entertainment (TIDSE), Darmstadt, Germany (2003)
6. Montfort, N.: Twisty Little Passages: An Approach to Interactive Fiction. MIT Press, Cambridge (2003)
7. Packard, E.: The Cave of Time. Bantum Books, New York (1979)
8. Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D.R., Swartout, W.: Toward a new generation of virtual humans for interactive experiences. IEEE Intell. Syst. **17**, 32–38 (2002)
9. Roemmele, M., Bejan, C., Gordon, A.: Choice of plausible alternatives: an evaluation of commonsense causal reasoning. In: Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University (2011)
10. Roemmele, M., Gordon, A.S.: Creative help: a story writing assistant. In: Schoenau-Fog, H., Bruni, L.E., Louchart, S., Baceviciute, S. (eds.) ICIDS 2015. LNCS, vol. 9445, pp. 81–92. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27036-4_8
11. Swanson, R., Gordon, A.S.: Say anything: using textual case-based reasoning to enable open-domain interactive storytelling. ACM Trans. Interact. Intell. Syst. **2**(3), 16:1–16:35 (2012)
12. Traum, D., et al.: New dimensions in testimony: digitally preserving a holocaust survivor's interactive storytelling. In: Schoenau-Fog, H., Bruni, L.E., Louchart, S., Baceviciute, S. (eds.) ICIDS 2015. LNCS, vol. 9445, pp. 269–281. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27036-4_26