Causality in Hundreds of Narratives of the Same Events

Emmett Tomai[1], Laxman Thapa[1], Andrew S. Gordon[2] and Sin-Hwa Kang[2]
[1]Department of Computer Science, University of Texas - Pan American, 1201 West University Dr.
Edinburg, TX 78539 USA
[2]Institute for Creative Technologies, University of Southern California, 12015 Waterfront Dr.
Los Angeles, CA 90094 USA
(etomai@cs.panam.edu, lthapa@broncs.utpa.edu, gordon@ict.usc.edu, kang@ict.usc.edu)

Abstract

Empirical research supporting computational models of narrative is often constrained by the lack of large-scale corpora with deep annotation. In this paper, we report on our annotation and analysis of a dataset of 283 individual narrations of the events in two short video clips. The utterances in the narrative transcripts were annotated to align with known events in the source videos, offering a unique opportunity to study the regularities and variations in the way that different people describe the exact same set of events. We identified the causal relationships between events in the two video clips, and investigated the role that causality plays in determining whether subjects will mention a particular story event and the likelihood that these events will be told in the order that they occurred in the original videos.

Analysis of Narrative

Causality plays a central role in narrative. In artificial intelligence research, logical formalisms of action and planning have long been shown to provide effective ways to represent and manipulate causal structures (cf. Wilensky 1983; Riedl and Young 2010). However, a computational model of narrative requires more than just causal soundness. There are complex and idiosyncratic psychological processes that dictate the presentation choices a narrator makes, and considerable research has been done on the impact of causal structure. In the field of discourse psychology, Trabasso and van den Broek (1985) developed the *causal network* model, a highly influential model of the causal structure of goal-based stories. Using this model, discourse psychologists have been able to predict a wide range of observed memory behaviors, sentence reading times, recognition priming latencies, lexical decision latencies, goodness of fit judgments for story sentences, and the inferences produced during thinking aloud (van den Broek 1995; Magliano 1999). This type of empirical research can inform better computational models. However, it is difficult to obtain suitable corpora for analysis. The narratives being studied must be paired with detailed descriptions of the situations that are being described. Further, variations in situations and variations in narrative style confound one another. Ideally, researchers would be able to analyze narratives from hundreds of subjects, where each is describing the same situation where the causal details are known.

We have identified an existing corpus of narratives, collected as part of a separate research effort, which comes close to this ideal. Gratch et al. (2007) describe a series of experiments to study the rapport that people can develop with interactive animated virtual characters. In that research, hundreds of human subjects told narratives describing the events of the same two video clips. The transcriptions of these verbal narrations (the Rapport Corpus) present a unique opportunity to study the factors that influence narrative choices. With hundreds of narratives of the same set of (known) events, it is possible to conduct quantitative analyses of the specific ways that the causal structure of the situation (the events of the two videos) plays a role in narrative discourse.

In this work, we have annotated the Rapport Corpus to link the narrative events in each transcript to the known events in the source videos. We have also identified the causal structure of those known events. Through a quantitative analysis of the resulting set of annotations, we explore the role that causality plays in determining two key aspects of narrative generation. First, we show that events with numerous causal connections to other events are more likely to be mentioned in a narration. Second, we show that the number of causal connections to and from an event impact the likelihood that that event is narrated out of the order in which it actually occurred.

The next sections of this paper describe the Rapport Corpus and our annotation process. We then present the methodology and results of our statistical analysis to answer our two research questions, and conclude with discussion.

The Rapport Corpus

This section describes the corpus of transcripts of verbal narrative that resulted from the work of Gratch et al. (2007), henceforth the *Rapport Corpus*. The original aim of that project was to assess the potential of an animated virtual character (the Rapport Agent) to create more engagement and speech fluency, as compared to a real human listener. The Rapport Agent attempts to exhibit human-like nonverbal feedback to facilitate rapport with a human speaker. Rapport of this sort is considered a key factor for

conflict resolution and negotiations, as well as improving psychotherapeutic effectiveness and test performance in classrooms. To generate these nonverbal behaviors, the Rapport Agent tracks the real human speaker's prosody, head movements and body posture in real time, and rapidly generates timely feedback using head nods and postural mirroring.

The original study design was a between-subjects experiment, where different conditions were based on variations of the algorithm used to generate the Rapport Agent's nonverbal feedback (Ning and Gratch 2009). During the experiment, subjects were first asked to watch two different video clips. One of the videos was a clip from a Tweety and Sylvester episode of Warner Bros.' *Looney Tunes* cartoon (henceforth, the *Tweety* video). The other video was the "CyberStalker" clip taken from a live-action segment from Edge Training Systems, Inc.'s Sexual Harassment Awareness video (henceforth, the *Sexual harassment* video). The order of the two video clips was randomized for each subject. Subjects were then asked to tell the story of the videos to the virtual agents. Subjects were told that the virtual agent was an avatar of a real human who was listening to their stories.

Subjects wore a headset so that their interaction with virtual agents could be recorded. Annotators later transcribed subjects' utterances from these recordings, and made annotations concerning their delivery. The annotation included codes for intonation, pause, pronunciation, laughter, and volume, from which researchers studied incomplete words, prolonged words, and pause fillers representing disfluencies in the subjects' storytelling.

The original study measured subjects' self-reported feeling of rapport, their verbal behaviors and their recall test scores on the stories in the video clips. The results demonstrated that the virtual agents' feedback immediacy elicited subjects' greater feelings of rapport. The overall duration of the subjects' verbal behaviors was longer when they interacted with a virtual agent that presented timely immediate feedback, as opposed to the other types of agents. Subjects talked longer when retelling the events of the second video that they viewed. However, the timely immediate feedback of the virtual agent did not facilitate improved recall performance.

The Rapport Corpus consists of 293 transcriptions of spoken narrative, 147 describing the events of the *Tweety* video and 146 describing the events of the *Sexual harassment* video. We obtained these transcripts directly from the authors of Gratch et al. (2007).

Annotation of Transcripts

In our project, the Rapport Corpus transcripts were annotated to connect the narrative descriptions to the ground truth of the videos being recounted. Each transcript was divided into utterances delineated by pauses of greater than 150 milliseconds. A sample excerpt from one of the transcripts follows, with numbered utterances.

```
1. okay um
2. it was a cartoon
3. and it
4. started out with a cat
5. um a black cat who was
6. at the
7. bird watchers society
8. and he was looking out the window with
   some binoculars and he
```

Each utterance was annotated with the events and details in the source video that it describes. The annotation guide contained a master list of events and details developed for each video. The first phase of development identified the (overlapping) sequence of directly observable physical events in each video. These events are quite precise and objective (e.g. "The cat turned his head from side to side"), but represent a low level of abstraction unlikely to be mentioned by a human narrator. Non-directly observable physical events that are highly implied by narrative conventions (e.g. a character zips by and there is the sound of an off-screen crash) were also included. The second phase addressed the level of abstraction issue. We took a small development set of transcripts from each video's set of transcripts. The events mentioned in those development transcripts were taken as representative of the level of abstraction for the entire set. Those events were added to the master list, subsuming the more fine-grained events that compose them. Those fine-grained events were then added to the new event as details as shown in Figure 1.

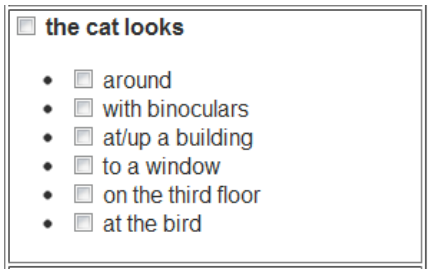

Figure 1. Event annotation example with event details.

As also shown in Figure 1, the master event list uses a semi-structured representation of events. As a guide for the annotation process, it was necessary that it could be read and understood by the annotators. Each event in the master is described as a SUBJECT-VERB phrase. Additional details are included as separate clauses, either direct object, prepositional phrase or directly quoted dialogue. In the case of simultaneous action, a detail clause is used to describe the additional verb in the same manner as the event (e.g. "and says"). The format is not hierarchical, leaving some ambiguity in the attachment of subsequent details, which better fits the level of abstraction the annotators were expected to work at.

Each master event list reflects the temporal sequence in the corresponding source video. Thus, annotation references not only the event as a type, but also the part of the timeline

in which it occurs. This fits well with the sequential nature of narrative. However, the videos also present situational details, such as where the characters are or what is visible behind them. These details are temporally bound to a situational interval at a higher level of abstraction than a narrator is likely to direction mention. Thus, these details are not tied to any of the events in the master. In practice, narrators mention them at any time during the presentation of events where their interval holds. This proved difficult for annotation. With events, moving back in the master to a previous event means that the narrator executed some form of flashback or sequencing error. This is not true for situational details. To alleviate potential confusion, the master list is divided into conceptually separate sections. The first shows the events, in sequence. The second shows situational details, which are not tied to that sequence. The third contains additional codes.

Acknowledging the degree of subjectivity in the annotation task, the annotation guide also contains codes to allow annotators to indicate utterances that do not strictly match the master event list. These codes were added to collect additional, more subjective data, but also to give the annotators various "none of the above" type options for difficult utterances. The intention was to relieve pressure on the annotators to make things fit, helping keep the event and detail annotations simple and accurate. The first set of codes cover summary statements, such as story structure (e.g. "his first attempt..."), genre commentary (e.g. "every Tweety cartoon", "like all training videos") and emotional impressions about atmosphere, quality and value judgments. The second set of codes cover character assumptions such as appearance (e.g. race, physical characteristics), thoughts, motives, feelings and character traits. The third deals with accuracy, allowing annotation of events and details missing from the master. The annotations for missing events and details are accurate, assumed, inaccurate and hypothetical. The final code is for observations made by the narrator that are not about the video at all (e.g. "I can't remember...").

The four annotators were undergraduate computer science and engineering majors at the University of Texas – Pan American. They were hired for the summer specifically to complete this annotation. The annotators each did several training runs on the development sets. They would independently annotate two or three transcripts, then meet with a project leader to provide insight on the decisions they made and go over the guidelines again. The early training runs, involving two of the annotators, were also a significant part of the development of the *Tweety* master event list. The *Sexual harassment* master event list (done second), although very different and challenging due to the heavy use of dialogue, did not require that iterative development.

Each annotator was given a period of several weeks to complete each of the two transcript sets. They worked using a web interface that we developed for this task. For each set, they were instructed to first watch the source video. Then, they logged in to the tool where the transcripts were provided in random order (within each set). They were instructed to first randomly read 10 of the transcripts, then work through the set one at a time. For each transcript, the annotator first went to a page where they read through the complete transcript in plain text, with one line per utterance and special annotation characters removed. He or she then moved to the annotation page for that transcript. On this page, the transcript was displayed line by line on the left side. On the right side, the annotations were displayed as shown in Figure 2.

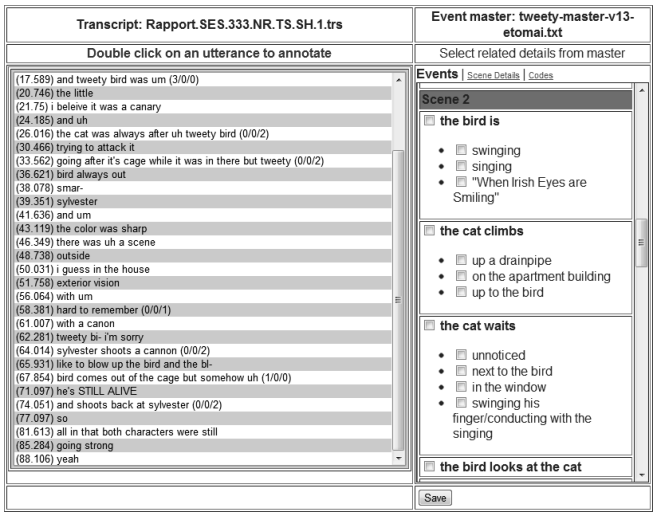

Figure 2. The web-based annotation tool.

They were instructed to go through the transcript line by line following these rules:

1. Double click on the utterance to select it (and save the last one)
2. If the utterance is information free (e.g. "um"), skip it
3. If the utterance is a connecting part of a bigger phrase (e.g. "where", "to", "at", "and then he"), and adds no details, skip it
4. Find and select the most appropriate events in the master that the utterance is describing
   a. If there is no appropriate event, leave it blank
   b. If you selected an event, select any details in that event that the utterance describes
   c. The same events or details can be selected multiple times, people tend to say redundant things
5. Select any additional scene details given in the utterance that aren't covered by the events
6. Select any of the codes that pertain to the utterance

The *Tweety* data set consisted of 147 transcripts, averaging 91 utterances per transcript. Five were used for development and removed. The remaining 142 were annotated by each of the 4 annotators. Out of 36 events present in the master event list, an average of 15 unique events were annotated in each transcript. Inter-rater agreement was calculated pair-wise using Cohen's kappa

(Cohen 1968). For the six pairings, the average kappa was 0.894 (std dev. 0.021).

The *Sexual harassment* data set consisted of 146 transcripts, averaging 94 utterances per transcript. Five were used in development and removed,. The remaining 141 were annotated by the same 4 annotators. Out of 38 events present in the master event list, an average of 18 unique events were annotated in each transcript. Inter-rater agreement was calculated in the same way. For the six pairings, the average kappa was 0.769 (std dev 0.018).

To facilitate analysis, the annotation data was post-processed and loaded into a different relational schema. Because the utterance boundaries are defined by pauses, and do not directly represent grammatical or conceptual boundaries, most events span multiple utterances. Many of those utterances are disfluencies, backtracking, or even unfinished sentences. Rather than attempt to annotate the span of each event, the annotators were tasked only to identify the events and details being clearly described in any given utterance. To facilitate analysis of event inclusion and ordering, the annotation data was post-processed to convert the annotated events from the per-utterance link representation to an utterance span representation. Each event is considered to begin at the first utterance it is annotated, and end at the last utterance it is annotated. To account for redundant mentions, events are allowed to overlap only 1 utterance. Thus, if one event is annotated from utterance 1 to 10, and another at utterance 5, the first event is divided into two separate events.

An additional set of unified annotations was created to reflect events agreed on by the group of annotators. Annotators differed on which specific utterances described an event, but generally agreed as to whether an event was described at all. The unified event annotation set was created by simply taking the union of the overlapping event spans, discarding initial and trailing intervals identified by only one annotator.

Experimental Methods

The availability of an annotated narrative corpus of this size affords opportunities for a wide range of analyses. In this work, we focused on two questions regarding the impact of causality on natural storytelling decisions. These questions required us to identify the causal relationships that exist among the events in the two situations (the *Tweety* and *Sexual harassment* videos). For this task, we followed the format used by Trabasso and van den Broek (1985) in their *causal network* model, where identified events are assigned causal links. For the identified events, we used the events listed in the two master guides that were used by our transcript annotators. For each master event, we listed each of the other events that could be viewed as their causal antecedents and consequents, producing a directed causal graph of the events in each video. Figure 3 depicts the causal graph for the events in the *Tweety* video, where arrows lead from a causal antecedent to its consequent. A similar causal graph was created for the events in the *Sexual harassment* video.

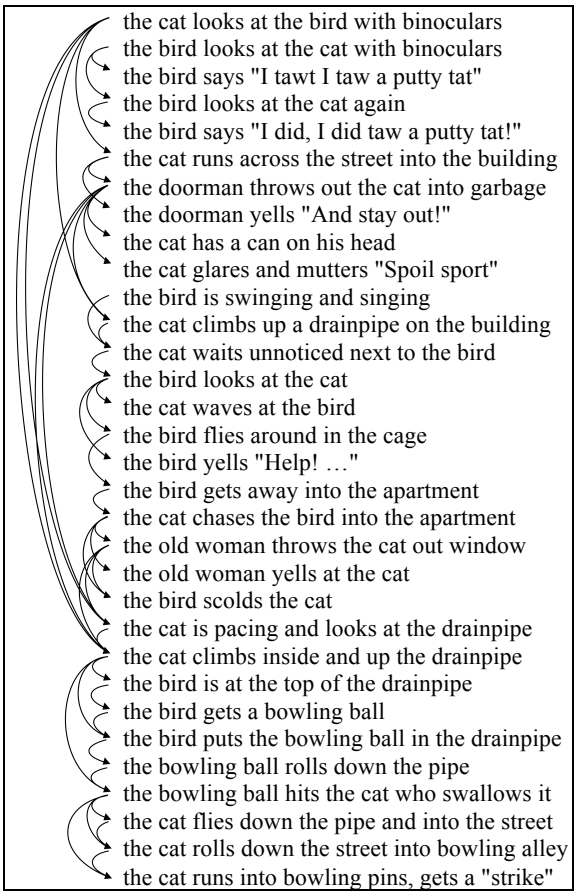

Figure 3. Causal analysis of the events in the *Tweety* video. Master event descriptions have been edited here for brevity.

As Figure 3 illustrates, events can be distinguished by the quantity of causal links leading to and from them. Some events have low causal connectivity, e.g. "the cat has a can on his head" has only one connection, as the consequent of "the doorman throws out the cat into garbage." Conversely, some events have high causal connectivity, e.g. "the cat climbs inside and up the drainpipe" is the consequent of four events and is the antecedent of three events, for a total of seven causal connections. We tabulated the number of causal antecedents and consequents for each event in both causal graphs, which we subsequently used to answer our two primary research questions.

First, we were interested in answering the question: *Are events with many causal connections more or less likely to be mentioned by subjects than events with few causal connections?* To answer this question, we first sorted all the master events from both videos into one of three categories based on the number of causal connections the events had in our causal graphs. We tabulated the number of antecedents and consequents of each event separately, sorting them into

type	video	none		one		multiple		F	η^2	p
		M	*SD*	*M*	*SD*	*M*	*SD*			
antecedents	*Tweety*	31.68	15.60	28.12	15.14	60.05	19.01	368.42[a]	.73	<.001
	Sexual harrassment	24.05	15.00	41.33◊	20.29	43.81◊	16.62	119.88	.46	<.001
consequents	*Tweety*	12.33	10.19	42.66	19.86	64.22	21.62	602.98[a]	.81	<.001
	Sexual harrassment	16.07	11.93	46.58◊	17.63	46.78◊	22.61	318.47[a]	.70	<.001

Table 1. Effect of the number (percentage) of antecedents and consequents on event mentions. Covariance differences between each pair of conditions are significant (p<.05) using the Bonferroni correction except where indicated (◊).

type	Video	none		one		multiple		F	η^2	p
		M	*SD*	*M*	*SD*	*M*	*SD*			
antecedents	*Tweety*	6.97	10.19	1.06◊	2.52	2.28◊	5.56	30.60[a]	.18	<.001
	Sexual harrassment	2.86◊	6.61	8.57	7.58	4.46◊	5.60	30.89[a]	.18	<.001
consequents	*Tweety*	.71	2.23	2.62	4.09	4.49	6.90	24.85[a]	.15	<.001
	Sexual harrassment	.18	1.49	7.81◊	7.26	8.00◊	9.07	63.63[a]	.31	<.001

Table 2. Effect of the number (percentage) of antecedents and consequents on event reordering. Covariance differences between each pair of conditions are significant (p<.05) using the Bonferroni correction except where indicated (◊).

categories for zero links, exactly one link, or multiple links. For example, the master event "the cat waits unnoticed next to the bird" in Figure 3 was sorted into the categories for multiple antecedents and exactly one consequent. We then tabulated the percentage of events from each category that are mentioned in each of the transcripts. For example, a particular transcript may mention 60% of the events that have multiple antecedents. Table 1 lists the mean (M) and standard deviation (SD) of these percentages for each video and causal direction.

The second question that we explored concerned the choices that storytellers make about the order of the events in their narrations. It is common that the narrated order of events is different from the order the storyteller experienced them. The reasons for this are varied. Storytellers may have misremembered the actual order. They may choose to foreshadow certain events for narrative effect. They may employ grammatical constructions that require a switch in event order, e.g. "*then x happened, because y had just happened.*" We were interested to see if causality played a role in these narrative decisions: *Are events with many causal connections more or less likely to be narrated out of the order in which they were experienced?*

To answer this question, we developed a simple rule for identifying reordered events in the annotated corpus. A narrated event is *reordered* in a given transcript if the narrated event that precedes it in the transcript appears later in the master event list. For simplicity, the very first event mentioned in any transcript is treated as correctly ordered, regardless of what follows it. We again sorted the master events into three categories for no, exactly one, and multiple causal links, for both antecedents and consequents. For each transcript, we then counted the number of reordered events from each category, normalized by total number of events in the category. Table 2 presents the mean (M) and standard deviation (SD) of these percentages for each video and causal direction.

To determine whether the differences were significant, we analyzed the data using the Repeated Measures ANOVA, a variation of ANOVA for use when the same participants take part in all conditions of an experiment (Tabachnick and Fidell 2001). Because we did multiple comparisons for repeated measures variables, we further did post-hoc comparisons using the Bonferroni correction to produce the pair-wise cell difference. Mauchly's test indicated that the assumption of sphericity was violated ($\varepsilon > .75$) in several of these experiments, so the degrees of freedom were corrected in these cases using Huynh-Feldt[a] estimates. The results of these analyses are also given in Tables 1 and 2.

Results

In the first analysis, the results show that there was a significant effect of the number of antecedents on subjects' mentioning the event in their narration, both for the *Tweety* video ($F(1.86, 259.75) = 368.42$, $p < .001$) and for the *Sexual harassment* video ($F(2, 278) = 119.88$, $p < .001$). There was also a significant effect of the number of consequents on subjects' mentioning the event in their narration, both for the *Tweety* video ($F(1.99, 278.67) = 602.98$, $p < .001$) and for the *Sexual harassment* video ($F(1.91, 265.87) = 318.47$, $p < .001$). In each case, events with no antecedents or no consequents were significantly less likely to be mentioned than events with one or more.

In the second analysis, the results show that there was a significant effect of the number of antecedents on subjects' reordering the events in their narration, both for the *Tweety* video ($F(1.49, 208.73) = 30.60$, $p < .001$) and for the *Sexual harassment* video ($F(1.99, 277.41) = 30.89$, $p < .001$). There

was also a significant effect of the number of consequents on subjects' reordering the events in their narration, both for the *Tweety* video ($F(1.56, 218.83) = 24.85$, $p < .001$) and for the *Sexual harassment* video ($F(1.63, 226.96) = 63.63$, $p < .001$). However, the sorts of events that were most likely to be reordered differed between the two videos. In the *Tweety* video, events with *no* antecedents were more likely to be narrated in an order different than which they were experienced. In the *Sexual harassment* video, events with exactly one antecedent were most likely to be reordered. With respect to the consequents, both videos exhibited the same pattern: events with no consequents were less likely to be narrated in an order different than which they were experienced.

Discussion

Our first finding is not surprising, either from an intuitive point of view, or in light of the body of research on the centrality of causality in recall. The main contribution of that finding lies in the qualities of the corpus. It is based on two highly diverse, natural artifacts, and consists of a large number of example narrations. In this study we have not only the conclusion that events mentioned in a narrative retelling are more likely to be causally significant, but also the annotated context in which that conclusion is drawn. It is our belief that the diversity of transcripts has much more to say about the construction of narrative retelling, particularly the identification of common versus individual preferences. The significance of causal events is a necessary backbone to further investigation.

Our second finding is less intuitively obvious. Considering the *Tweety* video first, we see that events with no causal antecedents are more likely to be reordered. This provides evidence that not only are causally significant events more memorable and important (to the narrator), but also that their position in the causal chain tends to be maintained. However, events with no causal *consequents* are *less* likely to be reordered. In the case where the event in question is at the end of a chain, this is consistent – a narrator is unlikely to reorder the final outcome if they are relying on those chains. The alternative case, where an event is completely causally unconnected, does not occur in the *Tweety* transcripts.

Considering the *Sexual harassment* video, there are a number of more complex factors. This video features a great deal of dialogue, much of which is recounting earlier events. This creates two different, overlapping timelines in the story itself: the conversation being observed, and the prior-to-ongoing events being discussed. As a result, the events that start causal chains in the observable conversation tend to have a single antecedent from the prior events being discussed. In contrast, the interactions in *Tweety* all begin with unexplained character action. This may explain why it was the single antecedent events in *Sexual harassment* that were more likely to be reordered. Certainly, the complexity of reported events mixed with observable events is an area for further investigation. The contrast between the *Tweety* and *Sexual harassment* videos in this study provides a useful first step in this direction.

Conclusion

In this paper we have presented a corpus of almost 300 narratives describing two short video clips. We annotated this corpus to link utterances in the narrations to unambiguous events and details in the sources, as well as links to more subjective concerns of character, atmosphere and narrator commentary. Using this annotated corpus, we were able to perform qualitative analysis of the impact of causal structure in the source situation on two types of choices made during narrative generation: event inclusion and reordering. Even these relatively simple findings can inform computational models of narrative, and the methodology presented here is a step towards increased empirical validation of narrative theories.

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70 (4): 213–220.

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M and Louis-Philippe Morency. (2007). Can virtual humans be more engaging than real ones? In *Proceedings of the 12th International Conference on Human-Computer Interaction*, Beijing, China.

Magliano, J. (1999) Revealing inference processes dur-ing text comprehension. In S. Goldman, A. Graesser, and P. van den Broek (Eds) *Narrative Comprehension, Causality, and Coherence: Essays in Honor of Tom Trabasso*. Mahwah, NJ: Erlbaum.

Riedl, M. and Young, R.M. (2010). Narrative Planning: Balancing Plot and Character, *Journal of Artificial Intelligence Research*, vol. 39.

Tabachnick, B. and Fidell, L. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Trabasso, T. and van den Broek, P. (1985) Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24: 612-630.

van den Broek, P. (1995) Comprehension and memory of narrative text: Inference and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539-588). New York: Academic Press.

Wang, N. and Gratch, J. (2009), Virtual Human Build Rapport and Promote Learning? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*.

Wilensky, R. (1983). *Planning and Understanding: A Computational Approach to Human Reasoning*. Reading, MA: Addison-Wesley.