Comparing Cognitive and Computational Models of Narrative Structure

David B. Christian and R. Michael Young

Department of Computer Science,
North Carolina State University, Campus Box 7535
Raleigh, NC, 27695
dbchrist@ncsu.edu, young@csc.ncsu.edu

Abstract

A growing number of applications seek to incorporate automatically generated narrative structure into interactive virtual environments. In this paper, we evaluate a representation for narrative structure generated by an automatic planning system by 1) mapping the plans that control plot into conceptual graphs used by QUEST, an existing framework for question-answering analysis that includes structures for modeling a reader's narrative comprehension and 2) using methods originally employed by QUEST's developers to determine if the plan structures can serve as effective models of the understanding that human users form after viewing corresponding stories played out within a virtual world. Results from our analysis are encouraging, though additional work is required to expand the plan language to cover a broader class of narrative structure.

Introduction

Detailed, immersive 3D virtual environments are increasingly commonplace in applications ranging from entertainment to social interaction to education and training. While many of these virtual environments include a narrative, the interaction within them is generally prescripted with little dynamic capability. Recent work (André, Rist, & Müller, 1998, Cavazza, Charles & Mead, 2002, Young & Riedl, 2003) seeks to develop algorithms that dynamically generate narratives for such environments. To date, however, there has been little effort to verify that the computational structures created to control interactive narratives correspond to mental models formed by people when comprehending activity within conventional story worlds (e.g., narrative texts).

We are currently developing *Mimesis*, a system that will automatically generate an internal description of the actions
composing a narrative's plot, then use that representation to
drive the action of characters within a 3D gaming environment (Young and Riedl, 2003). The action sequences used in Mimesis are automatically created by an artificial intelligence (AI) planning system called DPOCL

(Young, Pollack & Moore, 1994), described briefly in the following section. The research we describe here seeks to measure the correspondence between the data structures that Mimesis generates to represent plot and the mental models that people form when observing the action sequences as they are played out within the 3D story world.

To evaluate the correspondence between DPOCL plans and the mental models that users of our system form regarding an unfolding narrative, we first defined a mapping from DPOCL elements onto a subset of the conceptual graph structures that model narrative defined by Graesser, *et al,* in their work on QUEST, a psychological model of question answering (Graesser & Clark, 1985, Graesser & Franklin, 1990). Next, we generated a DPOCL plan capable of driving a story within our virtual environment and, using the mapping, determined the graph structure corresponding to the plan.

We then had human subjects view a video of the plan as it executed within a 3D virtual world, and asked them to rate the appropriateness of answers to questions regarding the story they had viewed. Their ratings were compared to QUEST's ratings of the same answers, a technique similar to that used in the evaluation of the QUEST algorithms themselves (Graesser, Lang, & Roberts, 1991). The results of the small-scale study are encouraging, indicating a strong correspondence between a number of features of both plans and the intended mental models users form.

Background

DPOCL Plan Structure

In this work, we represent the action that unfolds within a narrative-oriented virtual environment using a *plan*, a data structure completely characterizing the actions that occur in the story and the specification of the role that characters, objects and locations play in those actions. Plans and the planning systems that produce them have been widely studied in artificial intelligence research, though they are typically are used to describe specifications of real-world action sequences.

The plan structures we employ are those produced by the DPOCL planner (Young, Pollack and Moore, 1994).

Table 1: QUEST Knowledge Structure (QKS) Arc Types

Arc Type	Description (A -> B)	Narrative Structure Equivalent
Consequence	A causes or enables B	Actions have effects, which are state changes. Events cannot yet cause other events.
Implies	A implies B, semi-logical reasoning	Syllogistic reasoning is outside of scope, goal equivalency is handled through causal links from actions to goal states
Reason	B is a reason or motive for A B is a superordinate goal of A	Causal links – basic plan components – embody reasons for actions
Outcome	B specifies whether or not the goal A is achieved	In simple plan sctructures, goals are always achieved
Initiate	A initiates or triggers the goal in B	Not covered (plans for future coverage)
Manner	B specifies the manner in which A occurs	Not covered (plans for future coverage)

DPOCL plans are composed of *steps* corresponding to the actions that occur in a story; DPOCL uses the approach used by STRIPS (Fikes & Nilsson, 1971) to represent individual actions.

In DPOCL, each step is defined by a set of *preconditions*, the conditions in the world that must hold immediately prior to the step's execution in order for the step to succeed, and a set of *effects*, the conditions in the world that are altered by the successful execution of the action. In addition to a set of steps, a DPOCL plan contains a set of *temporal constraints* defining a partial temporal ordering on the execution of the plan's steps and a set of *causal links* connecting pairs of steps. Two steps *s1* and *s2* are connected by a causal link with associated condition *c* just when *c* is an effect of *s1* and a precondition of *s2* and the establishment of *c* by *s1* is used in the plan to ensure that *c* holds at *s2*. Further, DPOCL plans contain information about the hierarchical structure of a plan, similar to the representation used by hierarchical task network (HTN) planners (Sacerdotti, 1977). In DPOCL, each abstract step in a plan is connected to a set of more primitive steps that represent a sub-plan for achieving the abstract action's effects. Overall, DPOCL plans contain a fair amount of causal, hierarchical and temporal structure. This structure is added by the planning algorithm in order to guarantee that a plan is *sound*, that is, that, when executed, the plan is guaranteed to achieve it's goals. Recent work (Ratterman, et al, 2002, Young, 1999) suggests that this structure as well as the techniques used by the DPOCL algorithm to create it, make for effective models of human plan reasoning.

The QUEST Knowledge Structure

Graesser has developed a framework for representing the cognitive structures built by readers while reading texts (Graesser & Clark, 1985). Graesser's model is based on conceptual graph structures, which contain concept nodes and connective arcs. These graphs, referred to here as *QUEST knowledge structures* (QKSs), are used to describe the reader's conception of a narrative text and related general knowledge (Graesser & Clark, 1985). QKSs can be combined and traversed in order to make predictions relating to an idealized reader's conception of a text

(Graesser & Clark, 1985; Graesser & Franklin, 1990; Graesser & Hemphill, 1991; Graesser, Lang & Roberts, 1991).

A QKS consists of "statement" nodes connected by directed arcs. Both nodes and arcs are typed based on their meaning and purpose. Below is a short description of the statement nodes and relational arcs; the references above provide a complete discussion.

Statement nodes contain individual pieces of data within a QKS. Statement nodes contain either a simple sentence (e.g. of the form *<subject> <verb> <simple predicate>*) or any number of combinations of such sentences (e.g. "x saw P1" where P1 = *<subject> <verb> <simple predicate>*). These statement nodes are categorized into state, event, goal, style, and concept nodes. A *state* node describes a part of the world-state that can be assumed to remain true unless explicitly changed. An *event* node describes the changing of a state, with or without intention. *Goal* nodes describe the desire of an agent, although such goals do not have to be conscious. Any action taken intentionally has a corresponding goal and node.

Relational arcs connect statement nodes within the QKS, and by doing so describe how the narrative data represented within the QKS are cognitively linked. The six arc types that are crucial to event-oriented storytelling are displayed and described in Table 1 (1991). These arcs focus on events and the reasoning of characters, which constitute most of basic story telling.

Comparing Computational and Cognitive Structures

To determine whether or not DPOCL plans are cognitively plausible models of a user's understanding of a narrative controlled by those same plans, we developed a mapping from the DPOCL plan representation onto an existing, well-understood cognitive model, namely Graesser's QKS. We then tested whether a correspondence existed between the user's responses to questions about the plan-based QKS and the responses of the QUEST model. The mapping from DPOCL to QKS and the specifics of the experiment to determine the model's effectiveness are described below.

Let n be the number of top-level goals of plan P and m be the number of steps in P, then

1. Create a total ordering o for the m steps in P consistent with P's partial ordering over its steps.
2. For each of P's top-level goals g_i, $i = 1, .., n$, convert g_i into a goal node G_i.
3. For each plan step s_j in P, starting with the last step in o
    3.1 Convert s_j into a goal node G_j and an event node E_j
    3.2 Link G_j to E_j with an outcome arc
    3.3 For each effect e of s_j, add a state node S listing e and link E_j to S by a consequence arc
4. For each causal link in P connecting two steps s_1 and s_2 with condition c, connect the state node S created for s_1 and c in 3.3 to the event node E createad for step s_2 in 3.2
5. Number the nodes based on the total order o, using a depth-first algorithm for numbering sub-plan goals
6. For each causal link in P connecting two steps s_1 and s_2, find the goal nodes G_1 and G_2 created for s_1 and s_2 in 3.2. If G1 has a lower node number than G_2, connect G_1 to G_2 with a reason arc.

Figure 1: Mapping a plan structure to QUEST

Mapping a DPOCL Plan Structure to a QKS

The translation of plans to QKSs is not straightforward due to the expressivity of the QKS and the relative rigidity of plan structures. As noted, plan structures only have one type of arc (the causal link) and one type of node (the plan step) compared to Graesser's many. D espite these difficulties, it is possible to map a standard DPOCL plan structure into a simple yet functional QKS. Figure 1 describes the algorithm to perform this task. The resulting QKS is regularly structured, and is related in form and content to what Graesser calls a goal-oriented substructure (Graesser & Clark, 1985).

This mapping relies on the idea that many of Graesser's arc types have a simple corollary within DPOCL. Table 1 shows a listing of each arc type and its equivalent (or lack thereof) in a standard plan structure. This initial mapping includes only outcome, consequence, and reason arcs. The generated QKS is also limited to the actions of a single character, which makes for a relatively simplistic narrative. Future work will seek to expand the DPOCL plan representation and the mapping algorithm to address these limitations.

Experimental Evaluation

We wish to test whether the QKS generated by the automatic mapping from a DPOCL plan can be used as an effective means of predicting the "goodness of answers" (GOA) to why, how, and "what enabled" questions regarding the story that the plan describes. If QUEST's predictor variables perform as expected, it would be a strong indication that DPOCL plan structures can serve as an effective model of an idealized user's understanding of the narrative. The hypotheses and methods we use are modeled after Graesser's own tests of the QUEST algorithm's predictive ability (Graesser, Lang, & Roberts 1991): Participants watch a video, rate the goodness of answers (GOA) to questions about that video, and their responses are compared to the QUEST algorithm's predictions.

QUEST Predictor Variables

Goodness of answer (GOA) ratings are key to Graesser's validation of the QUEST model, and obtaining estimates of these ratings is the purpose of the QUEST predictor variables. There are three major predictor variables which are correlated to the goodness of an answer: an arc search procedure, the structural distance between the question and answer nodes, and constraint satisfaction. Their numeric values are calculated for a question/answer pair, and can then be used to predict human GOA ratings, or the likelihood that an answer will be produced when asked a given question (Graesser & Hemphill, 1991; Graesser, Lang, & Roberts, 1991; Golding, Graesser & Millis, 1990).

The *arc search procedure* differs depending on the question category. In all cases, certain types of arcs are traversed in certain directions starting from the question node. If a node can be reached via those arcs, it is considered to pass the arc search procedure, and receives a value of 1. Otherwise, it receives a value of 0.

The *structural distance* measures the number of arcs that must be traversed to reach the answer node from the question node. If a legal path is found using the appropriate arc search procedure, then the shortest legal path length is used. Otherwise, the shortest path length is used.

The *constraint satisfaction* measurement is an averaging of several variables including argument overlap, temporal compatibility, planning compatibility, plausibility, and causal strength. The value ranges from 0 to 1. For a complete description, see Graesser, Lang, & Roberts (1991).

Table 2: A Sample Training Session Guide

You are in a long hall, called the "lava room". Look around with the mouse to see what is in the lava room. In this training session, your job is to raise and then cross the bridge across the lava, which is now lowered. To raise and lower the bridge, you must use your hand on the hand symbol on the left-hand wall. The hand symbol is close to the lava on the same side as the power switch. The hand symbol for operating the bridge does not appear and the bridge does not work unless the power is on. You can turn on the power by using your hand on the yellow power switch. Cross the raised bridge to complete this training session.

Method

Participants. 15 computer science undergraduates from NC State University were given course extra credit participating in this study. The participants were randomly assigned into one of three subject groups. Each group was

asked to rate answers in one of three different categories: answers to "how", "why", or "what enabled" questions.

Materials. Participants first completed five training sessions within a virtual environment in order to familiarize themselves with the rules for action within the story world. Each training session required a subject to control a character via the mouse and keyboard. Participants were required to use their character to successfully perform a task composed of individual actions that also made up the story plans they would later observe being executed. Table 2 contains one training session guide, and Fig. 2 contains DPOCL step descriptions defined in the story world.

The 3D story world in which participants interacted modeled a four-room section of a spaceship. Rooms were connected by airlock doors, elevators and drawbridges with simple touch-activated controls. Actions in the story world typically had several preconditions and effects (e.g., the power must be on in order for a machine to process raw materials into medical supplies, pressing the drawbridge control button raises or lowers the bridge).

After the participants completed the five training sessions, they were shown a video that showed the execution of a DPOCL plan by a character acting in the same environment. The story involved the character processing raw materials into medical supplies and delivering them into another room. The video was derived from a plan with 20 actions; using our mapping algorithm this plan expanded into a 71 node QKS with 23 event/goal pairs and 25 state changes.

Name:	Name:
RaiseBridge(Bridge)	CrossBridgeEast(Bridge)
Preconditions:	**Preconditions:**
Lowered(Bridge)	Not(Lowered(Bridge))
PowerOn(Ship)	In(EastLavaRoom)
Effects:	**Effects:**
Raised(Bridge)	In(WestLavaRoom)

Figure 2: DPOCL Operators for the RaiseBridge and CrossBridge Actions

Seven event nodes and three state nodes from this QKS were chosen from the QKS as the basis for ten questions posed to the participants. The seven chosen event nodes were "cross the bridge," "raise the lift," "close the door from the medicine room to the hallway," "get the keycard," "turn on the power," "go from the lift to the elevated area," and "exit the lift into the medicine room." The three chosen state nodes were "have the medicine's ingredients," "the keycard door was open," and "the door from the hallway to the lava room was open." These nodes were turned into the appropriate questions by putting "Why did the protagonist," or "how did the protagonist," or "what enabled the protagonist to," etc., in front of the node's description.

For each question node, 16 answer nodes were chosen, leading to 160 question/answer pairs. In Graesser's

experiment, he used answers previously given to related questions in a previous experiment as a pool from which to draw his answer nodes. However, we did not have such a resource. Because Graesser's earlier experiments showed that the structural distance between the question and answer nodes was a key factor in determining the goodness of an answer and it was a straightforward measure, we picked at random eight nodes with a structural distance of three or less, and eight other nodes from throughout QKS.

Similar to Graesser's study, nodes that were part of a goal/event pair were described as a goal (e.g. "because she *wanted* to open the door"), if the event described in the answer occurs after the event described in the question node and as an event (e.g. "because she *opened* the door) otherwise. Answer nodes that described state changes were described as goals (e.g., "because she wanted the door to be open") if they had not happened yet, and as states ("because the door was open") otherwise. These descriptions were chosen order to give every action in the story its best chance of being rated as a good answer. The answer nodes were described in the same was for every phrasing of question, with the exception that "Because" was placed in front of the answers to why questions. Twelve answer nodes were thrown out because their wording did not follow these rules exactly, meaning that 148 question/answer pairs are used for all further analysis.

Procedure. The participants proceeded through the training sessions at their own pace. When they completed all five training sessions, they were told that they would be shown a video that took place in a similar world to the one they had just been working in. They were given the following back story, in order to express the initial goals of the narrative: "The video shows a woman who wants to bring some medicine from the medicine room to the small final room on the other side of the keycard door." They were told they could watch the video as many times as they wanted to, and that they should watch it carefully as they would be asked questions about it afterwards.

After the participants were finished with the video, it was explained that they would be rating answers to questions about the video they had just seen. The difference between good answers and bad answers was explained as described in Graesser, Lang, & Roberts (1991). However, no practice items were given. Participants rated the answers on a fourpoint scale, where answers could be a) bad answer to the question, b) possibly an acceptable answer, c) moderately good answer, or d) very good answer. Participants were timed as in Graesser's studies (Graesser, Lang, & Roberts, 1991) in order to preserve the format of the experiment, but those numbers were not used in our analysis.

Predictor Variables. Graesser used seven predictor variables with QUEST in order to examine whether the three main components of the QUEST algorithm were significant predictors of goodness of answer (GOA) ratings. However, the four auxiliary predictor variables are not relevant or calculable for this study. Therefore we

examined only the three central components of QUEST: the arc search procedure, structural distance, and constraint satisfaction. These components are described below.

Arc search procedure. Answers receive a 1 if there is at least one legal path to an answer using Graesser's arc search procedures for goal hierarchies, and a 0 otherwise. Because state nodes play a larger role in plan structures converted to QKSs, subordinate state changes were included in the arc search procedure wherever Graesser included subordinate actions. The mean arc-search procedure ratings were .22, .12, and .57 for why-, how- and enable- questions respectively (SD=.41, .32, .49).

Structural Distance denotes the number of arcs between nodes in the QKS. The shortest path found using the arc-search procedure was used if one existed; otherwise the shortest path was used. Structural distance was predicted to be negatively correlated with GOA scores. The mean structural distance score was 3.75 (SD=1.50).

Constraint Satisfaction is a complex variable that measures the argument overlap, temporal compatibility, planning compatibility, plausibility, and causal strength of an answer node. For a complete description, see Graesser, Lang, & Roberts (1991). Graesser had experts calculate these values; we estimated them computationally. These estimations were possible given the extra information about relationships contained within the plan structure. Argument overlap (AO) was derived from plan structures: if the a subset of the objects listed in the plan action for the question node where in the action for the answer node, then the answer received a 1 for this value. Temporal (TC) compatibility received a 1 if the answer was a goal node with a lower number, a 1 if the answer was a state change node that had not been negated by later state changes, a .5 if the answer were the directly previous or next action and a 0 otherwise. Planning compatibility and plausibility were assumed to be 1 and left out of final calculations because we involve only a single actor, and all of our answer nodes are considered plausible.

Causal strength was the combination of four factors: temporality, operativity, necessity, and sufficiency. Temporality (T) was 1 if the answer node had a lower number than the question node. Operativity (O) was considered equivalent to temporal compatibility above. Necessity (N) was 1 if the answer node was a state node and fulfilled a precondition for the question's related action, or if the answer node was an action and some effect of the action was such a state node. Sufficiency (S) was 1 if the answer was a goal node for which a precondition would be filled by taking the action listed in the question.

These numbers were combined to create constraint satisfaction score between 1 and 0 as follows: $(AO + TC. + (T \times O \times ((N + S) / 2)))/3$. The mean constraint satisfaction score was .43 (SD=.20). In general, the mean scores for all three predictor variables were not as high in our study as in Graesser's (by QUEST's criteria, our answers were poorer). This is most likely due to the different method used to select answer nodes.

Results

A multiple regression analysis was run to determine the effects of the three predictor variables, and a second test was run to determine the effects of interactions between the three variables. Graesser's earlier analyses predict that the regression equations will show a significant positive coefficient for arc-search procedures and constraint satisfaction and a significant negative coefficient for structural distance. Graesser's analyses also predict a significant interaction between arc-search and structural distance as well as between constraint satisfaction and structural distance.

Table 3 shows the standardized regression coefficients (β-weights) for each predictor variable. The table shows that arc-search and structural distance are important factors for all three question types. Structural distance is shown to be important only for enable questions.

Table 3: β-Weights of Predictors of Goodness of Answer (GOA), Segregated by Question Category

Predictor variable	Why	How	Enable
Arc Search	$.40_s$	$.69_s$	$.22_i$
Structural Distance	-.11	-.02	$-.25_s$
Constraint Satisfaction	$.34_s$	$.30_s$	$.28_s$
R^2	.41	.64	.29

Note: s = significant at $p < .05$ for question/answer pair.

We performed a second multiple regression analysis with the four possible interactions between the predictor variables added. A possible interaction was found between arc search and structural distance for enable questions $F(1,148) = 16$, $p < .06$. Adding in this interaction increased the variance explained (R2) from .29 to .31. No other interaction was found between the predictor variables.

Mean GOA ratings for why-, how- and enable- questions were calculated to check for ceiling or floor effects. The mean ratings for why-, how- and enables- questions were 1.66, 1.8, and 1.4 respectively, with standard deviations of .76, .88, and .62. These GOA scores are slightly more restrictive than Graesser's, who had means 1.99, 2.07, and 2.32 with standard deviations of .87, .80, and .85. This restricted range is likely due to the different method we used to select answer nodes, which was described in the materials section.

Discussion

In this study, we examined how well Graesser's QUEST model worked when the knowledge structure on which the algorithm was applied was automatically generated from a plan structure, all predictor variables were computed without human input, and the narrative was shown in a video instead of given in text. Despite the deviation from

Graesser's original design, two out of three of Graesser's QUEST components (arc search and constraint satisfaction) proved to significantly predict goodness of answers (GOA) ($p < .05$).

Structural distance was not shown to be a significant predictor for how and why questions, while in Graesser's study it was shown to be inversely related to GOA for all three question types. This discrepancy could be due to several factors. First, Graesser's β-weights for structural distance were small (-.16, -.13, -.13 for why-, how- and enable- questions); it could be that our sample size (N=15, N=5 for each question type) is too small to measure this effect. Second, it could be that Graesser's distance measures were correlated to the way answer nodes were chosen, and that this effect did not appear in our analysis due to our different method. Finally, our QKS covers only one actor's plan. Structural distance could only be a factor across multiple plans or between different actors.

No significant interactions were found between the three predictor variables ($p < .05$). It should be noted that the interactions found in Graesser's work all involved distance -- not a statistically significant predictor variable in two out of three question types. Thus, it is not surprising that there should be no significant interactions for these question types. In contrast, structural distance was found to be a statistically significant predictor in enable questions, and there is a possible interaction between arc search and structural distance ($p < .06$). However, because of the nature of the interaction and the small sample size, we did not analyze this possibility further.

The variance explained by Graesser's QUEST components was smaller in our study than in Graesser's. Our average R2 was 44.6, whereas Graesser's was 57.6. The variance explained for enable questions was 29%, as opposed to Graesser's 55%. The same factors that made structural distance less important may also have affected the overall effectiveness of QUEST's algorithm.

Conclusions and Future Work

One long-term goal of our research is to define a cognitively plausible structure for representing a viewer's understanding of a story that a) can be generated automatically and b) is expressive enough to be used to control a virtual environment. The results provided here are encouraging; the data indicate that DPOCL plan structures can be mapped automatically onto a QUEST knowledge structure, and that the resulting QUEST components are satisfactory initial predictors of GOA ratings for why-, how- and enable- questions in their new environment.

Conclusions and Future Work

References

André, E., Rist, T., & Müller, J. (1998). Integrating reactive and scripted behaviors in a life-like presentation agent. In *Proceedings of the 2nd International Conference on Autonomous Agents* (pp. 261-268).

Cavazza, M., Charles, F. & Mead, S., (2002) Planning characters' behaviour in interactive storytelling. In *Journal of Visualization and Computer Animation 13*(2): 121-131.

Fikes, R., Nilsson, N., (1971) STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. In *the Proceedings of IJCAI*, pages 608-620.

Graesser, A.C., Clark, L.F., (1985). *Structures and procedures of implicit knowledge*. Norwood NJ: Ablex.

Graesser, A.C, Franklin, S. P. (1990). Quest: A cognitive model of question answering. *Disc. Processes*, 13,279-303.

Graesser, A. C., & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory and Language, 30*. 186-209.

Graesser, A.C., Lang, K. L. & Roberts, R.M., (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General, 120*, 254-277..

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psycological. Review, 101, v.3,* 371-395.

Rattermann, M. J., Spector, L., Grafman, J., Levin, H. and Harward, H. Partial and total-order planning: evidence from normal and prefrontally damaged populations, in *Cognitive Science*, vol. 25, issue 6, pages 941-975.

Sacerdoti, E. (1977) *A Structure for Plans and Behavior*. Elsevier, New York.

Young, R. M, Pollack, M., & Moore, J. (1994) Decomposition and causality in partial-order planning, in *Proceedings of the Second International Conference on AI and Planning Systems*, Chicago, IL, pages 188-193.

Young, R. M. & Riedl, M. (2003) Towards an Architecture for Intelligent Control of Narrative in Interactive Virtual Worlds, in the *Proceedings of IUI,* January, 2003.

Young, R. M. (1999) Using Grice's Maxim Of Quantity To Select The Content Of Plan Descriptions. *Artificial Intelligence*, no 115, 215-256