

Natural Language Generation Journeys to Interactive 3D Worlds*

Invited Talk Extended Abstract

James C. Lester and William H. Bares
Charles B. Callaway and Stuart G. Towns
Multimedia Laboratory
Department of Computer Science
North Carolina State University
Raleigh, NC 27695

{lester, whbares, cbcallaw, sgtowns}@eos.ncsu.edu

<http://multimedia.ncsu.edu/imedia/>

Abstract

Interactive 3D worlds offer an intriguing testbed for the natural language generation community. To complement interactive 3D worlds' rich visualizations, they require significant linguistic flexibility and communicative power. We explore the major functionalities and architectural implications of natural language generation for three key classes of interactive 3D worlds: *self-explaining 3D environments*, *habitable 3D learning environments*, and *interactive 3D narrative worlds*. These are illustrated with empirical investigations underway in our laboratory with several such systems.

Introduction

Natural language generation (NLG) has witnessed great strides over the past decade. Our theoretical underpinnings are firming up, our systems building activities are proceeding quickly, and we are beginning to see significant empirical results. As a result of this maturation, the field is now well positioned to attack the challenges posed by a new family of computing environments: *interactive 3D worlds*, which continuously render the activities playing out in rich 3D scenes in realtime. Because of these worlds' compelling visual properties and their promise of a high degree of multimodal interactivity, they will soon form the basis for applications ranging from learning environments for education and training to interactive fiction systems for entertainment.

Interactive 3D worlds offer an intriguing testbed for the NLG community for several reasons. They may portray scenes with complicated spatial relationships,

such as those found in the domain of electricity and magnetism in physics. They may include multiple dynamic objects tracing out complex motion paths, such as water particles traveling through xylem tissue in virtual plants. They might be inhabited by user-directed avatars that manipulate objects in the world and lifelike agents that will need to coordinate speech, gesture, and locomotion as they explain and demonstrate complex phenomena. In 3D interactive fiction systems, user-directed avatars and lifelike autonomous agents may navigate through complex cityscapes and interact with users and with one another to create new forms of theater.

As the visual complexities of interactive 3D worlds grow, they will place increasingly heavy demands on the visual channel. To complement their rich visualizations, interactive 3D worlds will require the linguistic flexibility and communicative power that only NLG can provide. In interactive learning environments, the spatial complexities and dynamic phenomena that characterize physical devices must be clearly explained. NLG delivered with speech synthesis will need to be carefully coordinated with 3D graphics generation to create interactive presentations that are both coherent and interesting. In a similar fashion, lifelike agents roaming around the same 3D worlds through which users guide their avatars will require sophisticated NLG capabilities, and 3D interactive fiction systems will benefit considerably from virtual narrators that are articulate and can generate interesting commentary in realtime.

In this talk, we will explore the major issues, functionalities, and architectural implications of natural language generation for interactive 3D worlds. Our discussion will examine NLG issues for three interesting classes of interactive 3D worlds:

- **Self-Explaining 3D Environments:** In response to users' questions, self-explaining environments dynamically generate spoken natural language and 3D animated visualizations and produce vivid explana-

* Support for this work was provided by the following organizations: the National Science Foundation under grants CDA-9720395 (Learning and Intelligent Systems Initiative) and IRI-9701503 (CAREER Award Program); the North Carolina State University IntelliMedia Initiative; the William S. Kenan Institute for Engineering, Technology and Science; and a corporate gift from Novell, Inc.

Figure 1: The PHYSVIZ Self-Explaining 3D Environment

tions of complex phenomena.

- **Habitable 3D Learning Environments:** In habitable learning environments, lifelike pedagogical agents generate advice combining speech and gesture as users solve problems by guiding avatars through 3D worlds and manipulating devices housed in the worlds.
- **Interactive 3D Narrative Worlds:** Virtual narrators generate fluid descriptions of lifelike characters' interaction with one another in response to incremental specifications produced by narrative planners and interactively-issued user directives.

To begin mapping out the very large and complex space of NLG phenomena in 3D interactive worlds, it is informative to examine the issues empirically. These issues are being studied in the context of several projects currently under development in our laboratory. First, self-explaining 3D environments must coordinate NLG with 3D graphics generation. These requirements will be discussed with regard to the PHYSVIZ (Townes, Callaway, & Lester 1998) and the PLANT-WORLD (Bares & Lester 1997) self-explaining 3D environments for the domains of physics and plant physiology, respectively. Second, in habitable 3D learning environments, lifelike agents must be able to generate clear language that is carefully coordinated with agents' gestures and movements as they interact with users in problem-solving episodes. We examine these

issues in the VIRTUAL COMPUTER (Bares *et al.* 1998; Bares, Zettlemoyer, & Lester 1998), a habitable 3D learning environment for the domain of introductory computer architecture. Third, virtual narrators for 3D interactive fiction should be able to generate compelling realtime descriptions of multiple characters' behaviors. These issues are illustrated with examples from the COPS&ROBBERS world (Bares, Grégoire, & Lester 1998), a 3D interactive fiction testbed.

In the talk, we discuss current efforts to introduce NLG capabilities into these worlds at several levels. This includes (1) discourse planning, as provided by the KNIGHT explanation planner (Lester & Porter 1997), (2) sentence construction, as provided by the the FARE sentence planner (Callaway & Lester 1995) and the REVISOR clause aggregator (Callaway & Lester 1997), and (3) surface generation, as provided by FUF (Elhadad 1991). Below we briefly summarize the requirements and issues of NLG for self-explaining 3D environments, habitable 3D learning environments, and interactive 3D narrative worlds. These will be discussed in some detail in the talk.

Generation in Self-Explaining 3D Environments

As graphics technologies reach ever higher levels of sophistication, knowledge-based learning environments and intelligent training systems can create increasingly

Figure 2: The PLANTWORLD Self-Explaining 3D Environment

effective educational experiences. A critical functionality required in many such systems is the ability to unambiguously communicate spatial knowledge. Learning environments for the basic sciences frequently focus on physical structures and the fundamental forces that act on them in the world, and training systems for technical domains often revolve around the structure and function of complex devices. Explanations of electromagnetism, for example, must effectively communicate the complex spatial relationships governing the directions and magnitudes of multiple vectors representing currents and electromagnetic fields, many of which are orthogonal to one another.

Because text-only explanations are inadequate for expressing complex spatial relationships and describing dynamic phenomena, realtime explanation generation combining natural language and 3D graphics could contribute significantly to a broad range of learning environments and training systems. This calls for a computational model of 3D multimodal explanation generation for complex spatial and dynamic phenomena. Unfortunately, planning the integrated creation of 3D animation and spatial/behavior linguistic utterances in realtime requires coordinating the visual presentation of 3D objects and generating appropriate referring expressions that accurately reflect the relative position, orientation, direction, and motion paths of the objects presented with respect to the virtual camera's view of

the scene.

To address this problem, we are developing the *visuo-linguistic explanation planning framework* for generating multimodal spatial and behavioral explanations combining 3D animation and speech that complement one another. Because 3D animation planners require spatial knowledge in a geometric form and natural language generators require spatial knowledge in a linguistic form, a realtime multimodal planner interposed between the visual and linguistic components serves as a mediator. This framework has been implemented in CINESPEAK, a multimodal generator consisting of a media-independent explanation planner, a visuo-linguistic mediator, a 3D animation planner, and a realtime natural language generator with a speech synthesizer. Experimentation with CINESPEAK is underway in conjunction with self-explaining environments that are being designed to produce language of spatial and dynamic phenomena:

- *Complex spatial explanations:* PHYSVIZ (Towns, Callaway, & Lester 1998) is a self-explaining 3D environment in the domain of physics that generates multimodal explanations of three dimensional electromagnetic fields, forces, and electric currents in realtime (Figure 1).
- *Complex dynamic behavior explanations:* PLANTWORLD (Bares & Lester 1997) is a self-explaining 3D environment in the domain of plant anatomy and

Figure 3: The VIRTUAL COMPUTER Habitable 3D Learning Environment

physiology that generates multimodal explanations of dynamic three dimensional physiological phenomena such as nutrient transport (Figure 2).

Generation in Habitable 3D Learning Environments

Engaging 3D learning environments in which users guide avatars through virtual worlds hold great promise for learner-centered education. By enabling users to participate in immersive experiences, 3D learning environments could help them come to develop accurate mental models of highly complex biological, electronic, or mechanical systems. In particular, 3D learning environments could permit learners to actively participate in the very systems about which they are learning and interact with lifelike agents that could effectively communicate the knowledge relevant to the user's task. For example, users could study computer architecture in a virtual computer where they might be advised by a lifelike agent about how to help a CPU carry data from RAM to the hard disk, or they could study the human immune system by helping a T-cell traverse a virtual lymph system. Properly designed, 3D learning environments that blur the distinction between education and entertainment could produce engrossing learning experiences that are intrinsically motivating and are solidly grounded in problem solving.

Lifelike agents that are to interact with users in

habitable 3D learning environments should be able to generate language that enables them to provide clear problem-solving advice. Rather than operating in isolation, generation decisions must be carefully coordinated with decisions about gesture, locomotion, and eventually prosody. In collaboration with the STEVE virtual environments tutor project at USC/ISI (Rickel & Johnson 1998), we have begun to design NLG techniques for *embodied explanation generation* in which the avatar/agent generates coordinated utterances (delivered with a speech synthesizer) and gestural and locomotive behaviors as it manipulates various devices in the world. Embodied explanation generation poses particularly interesting challenges in the following areas:

- *Deictic believability*: Lifelike agents must be able to employ referring expressions and gestures that together are both unambiguous and natural (Lester *et al.* 1998).
- *Socially motivated generation*: Lifelike agents must not only express concepts clearly but also create utterances that are properly situated in the current socio-linguistic context.
- *Embodied discourse planning*: Media allocation issues must be considered in adjudicating between expressing advice verbally or through agents' demonstrative actions.

Over the past two years, we have constructed a habitable learning environment for the domain of computer

Figure 4: The COPS&ROBBERS Interactive 3D Narrative World

architecture. The VIRTUAL COMPUTER (Bares *et al.* 1998; Bares, Zettlemoyer, & Lester 1998) (Figure 3) is a habitable 3D learning environment that teaches novices the fundamentals of computer architecture and system algorithms, e.g., the fetch-execute cycle. To learn the basics of computation, users direct an avatar in the form of a friendly robot courier as they execute instructions and transport data packets to appropriate locations in a 3D “town” whose buildings represent the CPU, RAM, and hard disk. We are beginning to investigate deictic believability, socially motivated generation, and embodied discourse planning in an lifelike agent that provides advice in the VIRTUAL COMPUTER.

Interactive 3D Narrative Worlds

While story generation has been an NLG goal that dates back more than a quarter century and text-based interactive fiction systems have been the subject of increasing attention, it is the prospect of coupling sophisticated NLG with 3D believable characters that offers the potential of achieving interactive fiction generation in a visually compelling environment. One can imagine different genres of 3D interactive fiction, many of which will involve a *virtual narrator* who comments on the events unfolding in the world. In much the same manner that sports announcers come in two varieties, play-by-play and color commentary, *virtual narrators* can provide both a descriptive account of the world’s

activities as well as a running analysis on their significance. To stress test NLG, we adopt three constraints on generation for 3D narrative worlds:¹

- *Realtime*: World events play out in realtime and can be modified by users. Consequently, the relevance of utterances is time-bound; generators must construct their utterances in realtime and cannot know in advance how the actions in the world will play out.
- *Non-interference*: Generators cannot themselves enact modifications on objects or characters in the world. As a result, they must cope with what they are dealt by world simulators and users’ actions.
- *Multiple, simultaneous events*: Multiple activities occur in the world at the same time. Consequently, generators must make time-bounded moment-by-moment content determination decisions that necessarily omit mention of many actions.

We have recently begun to study these issues in COPS&ROBBERS (Bares, Grégoire, & Lester 1998), a 3D interactive fiction testbed with multiple characters interacting with each other in an intricate cityscape. In COPS&ROBBERS (Figure 4), three autonomous characters, a policeman and two robbers, attempt to capture a lost money bag dropped by a careless bank teller. If

¹Elisabeth André and colleagues at DFKI are addressing similar issues in their realtime generator for the ROBOCUP competition.

the policeman finds the money bag first, he dutifully returns it to the bank, but if either of the two miscreants find the unclaimed money, they will scurry off to Joe's Bar to spend their new found loot. If the cop catches either robber carrying the money, he will immobilize him and return the money bag to the bank. When the narrative begins, the three characters meander randomly through the town searching for the lost money bag. At any time, users may affect the narrative by modifying characters' physical abilities such as their speed or eyesight.

Despite the relative simplicity of the testbed, it poses significant NLG challenges. Of particular interest are problems in the virtual narrator's expressing time sequence relations, concisely describing locations where particular events are occurring, and linking characters' actions to their intentions. Because events occur simultaneously, tense issues are problematic in accurately describing the temporal relations between events in sequential utterances. Especially difficult are generating precise disambiguating locative descriptions involving relative locations, direction of movement, and proximity of characters and structures in the world. Because it is often important to identify where a specific action has occurred, generators must be able to formulate locatives that are precise. Frequently, they must also be concise, because utterances that are too verbose will require excessive speaking times, causing the narration to miss other important events. Finally, generators must be able to communicate about characters' goals, actions, and the relation between the two. For example, if the cop is scurrying toward one of the robbers, rather than merely reporting the action, the generator should sometimes comment on the causal link between the cop's desire to obtain the money bag and his accosting the targeted robber.

A New Era for NLG

As a result of both technological and societal developments, the advent of a new era for NLG is upon us. On the technology front, high-end 3D graphics, as well as the 3D interactive worlds they will spawn, will make significant demands on NLG systems. On the societal front, we're beginning to see the rapid convergence of the software, telecommunications, and even the entertainment industries. This will undoubtedly provide significant impetus for integrating NLG into applications that could not have even been imagined at the inception of the field. With continued progress in theory, systems building, and empirical studies, we will be well positioned to meet the upcoming challenges.

Acknowledgements

Many people have contributed to the projects discussed in the talk. The authors would like to thank: the technical members of the the IntelliMedia Initiative's 3D team including Joël Grégoire, Ben Lee, Dennis Rodriguez, and Luke Zettlemoyer; the IntelliMedia ani-

mation and modeling team, led by Patrick FitzGerald, including Tim Buie, Mike Cuales, Rob Gray, and Alex Levy; Bruce Porter for his collaboration on the KNIGHT explanation system; Jeff Rickel for his collaboration on the pedagogical agents dialogue work; and especially Michael Elhadad for creating and generously assisting us with FUF for the past five years.

References

- Bares, W. H., and Lester, J. C. 1997. Realtime generation of customized 3D animated explanations for knowledge-based learning environments. In *AAAI-97: Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 347-354.
- Bares, W.; Zettlemoyer, L.; Rodriguez, D.; and Lester, J. 1998. Task-sensitive cinematography interfaces for interactive 3D learning environments. In *Proceedings of the Fourth International Conference on Intelligent User Interfaces*, 81-88.
- Bares, W.; Grégoire, J.; and Lester, J. 1998. Realtime constraint-based cinematography for complex interactive 3D worlds. In *Proceedings of the Tenth National Conference on Innovative Applications of Artificial Intelligence*.
- Bares, W.; Zettlemoyer, L.; and Lester, J. 1998. Habitable 3D learning environments for situated learning. In *Proceedings of the Fourth International Conference on Intelligent Tutoring Systems*. Forthcoming.
- Callaway, C., and Lester, J. 1995. Robust natural language generation from large-scale knowledge bases. In *Proceedings of the Fourth Bar-Ilan Symposium on the Foundations of Artificial Intelligence*, 96-105.
- Callaway, C. B., and Lester, J. C. 1997. Dynamically improving explanations: A revision-based approach to explanation generation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 952-58.
- Elhadad, M. 1991. FUF: The universal unifier user manual version 5.0. Technical Report CUCS-038-91, Department of Computer Science, Columbia University.
- Lester, J. C., and Porter, B. W. 1997. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics* 23(1):65-101.
- Lester, J.; Voerman, J.; Towns, S.; and Callaway, C. 1998. Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*. Forthcoming.
- Rickel, J., and Johnson, W. L. 1998. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*. Forthcoming.
- Towns, S. G.; Callaway, C. B.; and Lester, J. C. 1998. Generating coordinated natural language and 3D animations for complex spatial explanations. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*.