Generating Coordinated Natural Language and 3D Animations for Complex Spatial Explanations*

Stuart G. Towns and **Charles B. Callaway** and **James C. Lester**
Multimedia Laboratory
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-7534
{sgtowns, cbcallaw, lester}@eos.ncsu.edu

Abstract

Dynamically providing students with clear explanations of complex spatial concepts is critical for a broad range of knowledge-based educational and training systems. This calls for a realtime solution that can dynamically create 3D animated explanations that artfully integrate well-chosen speech with rich visualizations. Unfortunately, planning the integrated creation of 3D animation and spatial linguistic utterances in realtime requires coordinating the visual presentation of 3D objects and generating appropriate spatial phrases that accurately reflect the relative position, orientation, and direction of the objects presented. We present a visuo-linguistic framework for generating multimedia spatial 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 multimedia planner interposed between the visual and linguistic components can serve as a mediator. This framework has been implemented in CineSpeak, a multimedia explanation generator consisting of a visuo-linguistic mediator, a 3D animation planner, and a realtime natural language generator with a speech synthesizer. CineSpeak has been used in conjunction with a prototype 3D learning environment in the domain of physics to generate realtime multimedia explanations of three dimensional electromagnetic fields, forces, and electrical current.

Introduction

As multimedia technologies reach ever higher levels of sophistication, knowledge-based learning environments and intelligent training systems can create increasingly effective educational experiences. Moreover, if learning environments could leverage the growing body of work on intelligent multimedia systems in the form of knowledge-based 2D graphics generation (Roth, Mattis, & Mesnard 1991; Mittal *et al.* 1995), automated static 3D graphics production (Wahlster *et al.* 1993; Feiner 1985; Seligmann & Feiner 1991; Feiner & McKeown 1993), and 3D animation generation (Bares & Lester 1997; Butz & Krüger 1996; Christianson *et al.* 1996; Karp & Feiner 1993), they could fluently generate multimedia explanations that clearly communicate complex concepts. A critical functionality required in many domains 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 spatial explanations are notoriously inadequate for expressing complex spatial relationships, realtime multimedia spatial explanation generation could contribute significantly to a broad range of learning environments and training systems. This calls for a computational model of multimedia explanation generation for complex spatial knowledge. Unfortunately, planning the integrated creation of 3D animation and spatial linguistic utterances in realtime requires coordinating the visual presentation of 3D objects and generating appropriate spatial phrases that accurately reflect the relative position, orientation, and direction of the objects presented. Although a number of projects have studied the automated coordination of natural language and 2D graphics (Feiner & McKeown 1993), previous work on knowledge-based 3D animation either avoids accompanying narration altogether (Butz & Krüger 1996; Christianson *et al.* 1996; Karp & Feiner 1993), employs canned audio clips in conjunction with generated 3D graphics (Bares &

*Support for this work was provided by a grant from the NSF (Faculty Early Career Development Award IRI-9701503), the IntelliMedia Initiative of North Carolina State University, The William S. Kenan Institute for Engineering, Technology and Science, and an industrial gift from Novell, Inc.

Lester 1997), or focuses on either basic coordination issues (Wahlster *et al.* 1993) or on the challenges of incorporating animated characters (André & Rist 1996) rather than on coordinating the generation of language and visualizations for complex 3D spatial relationships.

To address this problem, we have developed the visuo-linguistic explanation planning framework for generating multimedia spatial 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 multimedia planner interposed between the visual and linguistic components serves as a mediator. This framework has been implemented in CINESPEAK, a multimedia explanation 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. CINESPEAK has been used in conjunction with PHYSVIZ (Figure 1), a prototype 3D learning environment in the domain of physics, to generate realtime multimedia explanations of three dimensional electromagnetic fields, forces, and electrical current.

Spatial Explanation Generation

A critical functionality of knowledge-based learning environments and training systems (Burton & Brown 1982; Hollan, Hutchins, & Weitzman 1987; Lesgold *et al.* 1992) is automatically providing students with clear explanations of spatial phenomena. For the same reason that in psycho-social frameworks of comprehension, hearers interpret linguistic events in concrete contexts, speakers (and hence, spatial explanation generators) must carefully consider the physical context in which utterances are generated (Fillmore 1975). Generating clear spatial explanations therefore entails addressing six fundamental problems, each of which can be illustrated with the difficulties presented by an explanation system for the domain of physics that must communicate the basic principles of electromagnetism:

- *Complementarity of 3D Animation and Speech:* Because of the conceptual complexity of spatial knowledge, even three-dimensional animations without accompanying explanatory speech are too limiting. For example, explanations of how to apply the right-hand rule[1] to solve E&M problems require both (1) spoken natural language about how the fingers and thumb correspond respectively to the current and magnetic force and (2) visual demonstrations of the spatial relationships bearing on the alignment of the thumb and fingers with the particular orientations of forces, fields, and current in the world. While previous work has addressed the coordination of 2D graphics and natural language (Maybury 1994; Feiner & McKeown 1993), work on 3D animation generation either does not address natural language generation issues (Bares & Lester 1997; Butz & Krüger 1996; Christianson *et al.* 1996; Karp & Feiner 1993) or does not explore natural language generation capabilities required of complex spatial knowledge (Wahlster *et al.* 1993; André & Rist 1996).

- *Physical Context Impact on Visuo-Linguistic Utterances:* Because of the inherent difficulties in linguistically expressing spatial relationships, generating spatial natural language poses enormous difficulties. While foundational work has studied generating spatial natural language, e.g., scene description generation (Novak 1987) and spatial layout description generation (Sibun 1992), the interplay between relative and absolute coordinate systems must be carefully monitored. For example, in explaining how magnets induce a field that flows from the north pole of a magnet to a south pole of another magnet, and explaining how current in a wire flows from positive electrodes to negative ones, the relative directions that the field and current travel in a physical environment depend on the absolute locations of the poles and electrodes. The language employed to realize this message is therefore highly dependent on (1) the orientation of objects in the world and (2) the students' perspective on these objects, e.g., whether she is viewing them from in front, to the side, or behind.

- *Synchronization of 3D Animation and Speech:* Just as in the coordination of natural language and 2D graphics (Maybury 1994; Feiner & McKeown 1993) when the timing of events must be considered, the timing of visual cues and events must be synchronized with the relevant spatial utterances for 3D. For example, in explaining how a particular section of a wire has a magnetic force acting on it, when the speech refers to that section of the wire, the animation might highlight that region when the reference to it is spoken.

- *Dual Representation of Geometric and Linguistic Spatial Knowledge:* While we are far from a comprehensive theory of spatial reasoning, which must included techniques for determining individuation, relative position, and relative orientation of objects (Davis 1990; Gapp 1994), integrated 3D spatial explanations combining animation with speech must exploit two types of representations of space. Animation planners for 3D visualizations reason most easily with geometric representations, while natural language generators require spatial representations that can enable them to map spatial relations to grammatically appropriate realizations.

[1] The *right hand rule* is a mnemonic device for determining the three dimensional orthogonal spatial relationships that hold between current, magnetic fields, and the resulting magnetic force they induce.

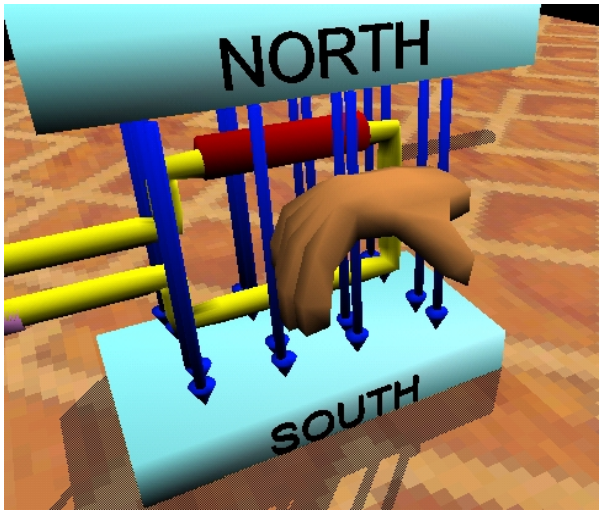

Figure 1: Explaining electromagnetism in the PHYSVIZ learning environment

Generating Coordinated 3D Spatial Explanations

As a student interacts with a 3D learning environment, they manipulate the 3D scene in which the the objects of interest are arranged. For example, a 3D learning environment for the domain of physics might include current-carrying wires and magnetic fields surrounding the poles of magnets. When the student poses a query (Figure 2), a media-independent explanation planner uses it to construct a plan for communicating that goal. By inspecting a knowledge base of domain concepts and using its explanation knowledge about how to communicate, it forms an explanation plan specifying the temporal order in which atomic presentation units should be conveyed. Critically, none of these specifications include low-level geometric or linguistic knowledge; they are restricted to references to domain objects and processes. A visuo-linguistic mediator examines the leaves of the plan and parcels out the specifications to a 3D animation planner and a natural language generator. To the animation planner, the mediator passes visual communicative goals that specify the objects that should be featured. It exploits knowledge of the scene geometries and the 3D models occupying the virtual world to create animation plans. To the language generator, the mediator passes linguistic communicative goals that specify the concepts to be realized in speech. It exploits a grammar capable of producing spatial utterances involving concepts related by direction and orientation and a lexicon with spatial entries to create the appropriate text.

To the great extent possible, the mediator requests both the animation planner and the language generator to run to completion. Because the animation planner makes determinations about the final positions of models, and hence the relative orientations of objects in visualizations, it can run undisturbed. However, because the language generator frequently requires up-to-date knowledge about the positions and orientations of the featured 3D models in order to generate appropriate spatial phrasings, it often must inform the mediator that its knowledge about spatial relationships is incompletely specified. The mediator consults the animation planner's world model and supplies the natural language generator with the necessary spatial features.

The 3D animation specifications and the natural language specifications of the explanation plans are passed to the media realization engines. The 3D animation specifications are passed to the animation renderer, while the text produced by the natural language generator is passed to a speech synthesizer. Visualizations and speech are synchronized in an incremental fashion and presented in atomic presentation units as dictated by the structure of the initial media-independent plan. They are presented in realtime within the 3D learning environment, and the process repeats each time the student poses another query.

Explanation Plan Construction and Visuo-Linguistic Mediation

Given a communicative goal to explain some complex spatial phenomenon, the media-independent explanation planner constructs an explanation plan that will be used in each of the upcoming phases. Using the by-now classic top-down decomposition approach to explanation generation (Suthers 1991; Cawsey 1992; Hovy 1993; Moore 1995), the media explanation determines the following:

- *Explanatory Content:* By extracting relevant propositions from the domain knowledge base, it identifies the key knowledge (spatial and otherwise) to include in the final explanation. For example, when a request to explain how the right-hand rule is used to determine the direction of the magnetic force acting on the wire, it then examines the knowledge base to find the inputs (current and magnetic field), the subevents (finger pointing and finger curling), and the outputs (the direction of the force).
- *Multimedia Rhetorical Structure:* It must then impose a temporal structure on the knowledge identified above. In the same manner that text has a discourse structure, multimedia explanations have an analogous structure that specifies the order in which to present content in the 3D animations and spoken utterances. For example, the content identified in the example above is organized in the structure depicted in the second level of the explanation plan shown in Figure 3.
- *3D Animation Specifications:* Each of the content specifications is annotated with visual presentation specifications. To maintain the high degree of modularity essential for such multi-faceted computations, it is critical that the media-independent explanation

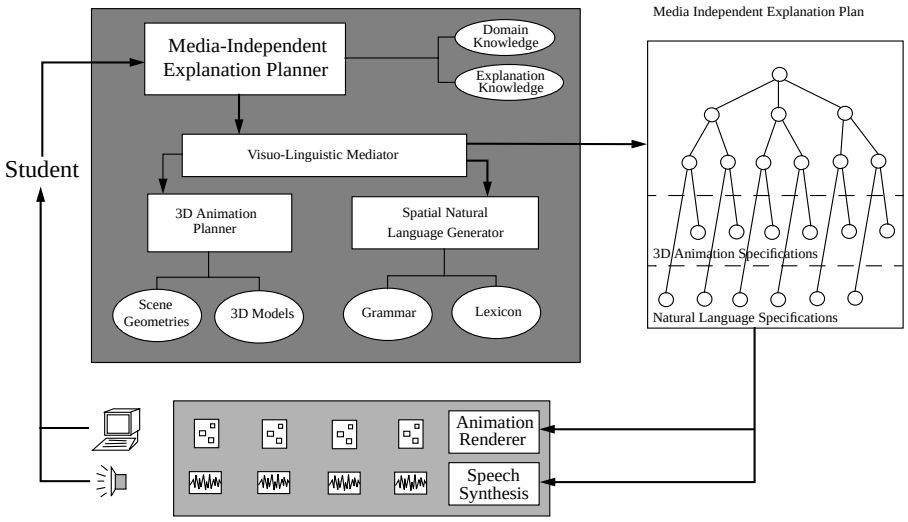

Figure 2: The Visuo-Linguistic Multimedia Explanation Framework

planner not be concerned with *any* of the complexities of 3D animation generation. To accomplish this, the explanation planner expresses its presentation needs with very high-level visual specifications. For example, to visually present the magnetic field, the explanation planner creates the visual annotation `(show-object magnetic-field)` to request that the animation planner create a clear shot of the magnetic field.

- *Linguistic Specifications:* Each of the content specifications of the explanation plan is also annotated with linguistic presentation specifications. As above, all details of natural language generation will be delegated to the linguistic component, so the explanation planner formulates the linguistic requirements without itself considering grammatical or lexicalizaton issues.

Once the media-independent explanation plan has been constructed, the visuo-linguistic mediator coordinates the integrated generation of visual and linguistic expressions of spatial knowledge in the content determined above. However, achieving the desired integration while preserving the modularity of the media planners is complicated by the fact that it (a) has no detailed knowledge of scene geometry and (b) has no detailed knowledge of linguistic techniques for realizing spatial knowledge in appropriate phrases. To address these problems, the mediator conducts itself as follows.

(1) The mediator issues recommendations to the natural language generator by formulating as much of a linguistic specification as it can. (2) If it encounters no spatial uncertainties, i.e., features in the evolving specifications with values that cannot be determined without detailed knowledge of scene geometries, its task is complete and no arbitration is required. Because of the dynamic nature of the virtual camera that "films" the animations, it is likely that spatial uncertainties will arise. For example, if the camera is filming a motor in the PhysViz environment from a front view, from the student's perspective, the current in the wire appears to flow to the left, so the utterance should communicate the notion of "leftward." In contrast, if the camera is filming exactly the same apparatus from a rear view, from the student's perspective, the current in the wire appears to flow to the right, so the utterance should communicate be express a "rightward" direction of flow. It is therefore the responsibility of the mediator to determine the correct orientations and inform the natural language generator. (3) To do so, on an as-needed basis, it requests spatial information from the animation planner, which computes spatial knowledge from scene geometries in its developing animation plan. (4) It next delivers the new spatial knowledge to the natural language generator. (5) Finally, it issues orders for both the animation planner and natural language generator to undertake their respective tasks.

3D Animation Planning

When the animation planner is invoked with high-level visual communication goals, its task is to construct a 3D visualization that clearly communicates the spatial concepts and relations. These include *positions* of objects, such as the north magnetic pole being on top of the motor, *orientations*, such as a magnetic field facing downwards, and *relative orientations*, such as the current in the wire being orthogonal to the magnetic field. Because animated explanations should focus students' attention on the most critical concept at each moment in the explanation (Rieber 1990), the animation planner must carefully lay out the low-level visual specifications which will be passed to the renderer.[2] Planning

[2]The 3D animation planner is the result of a long term effort to develop a general-purpose pedagogical 3D animation generator (Bares & Lester 1997).

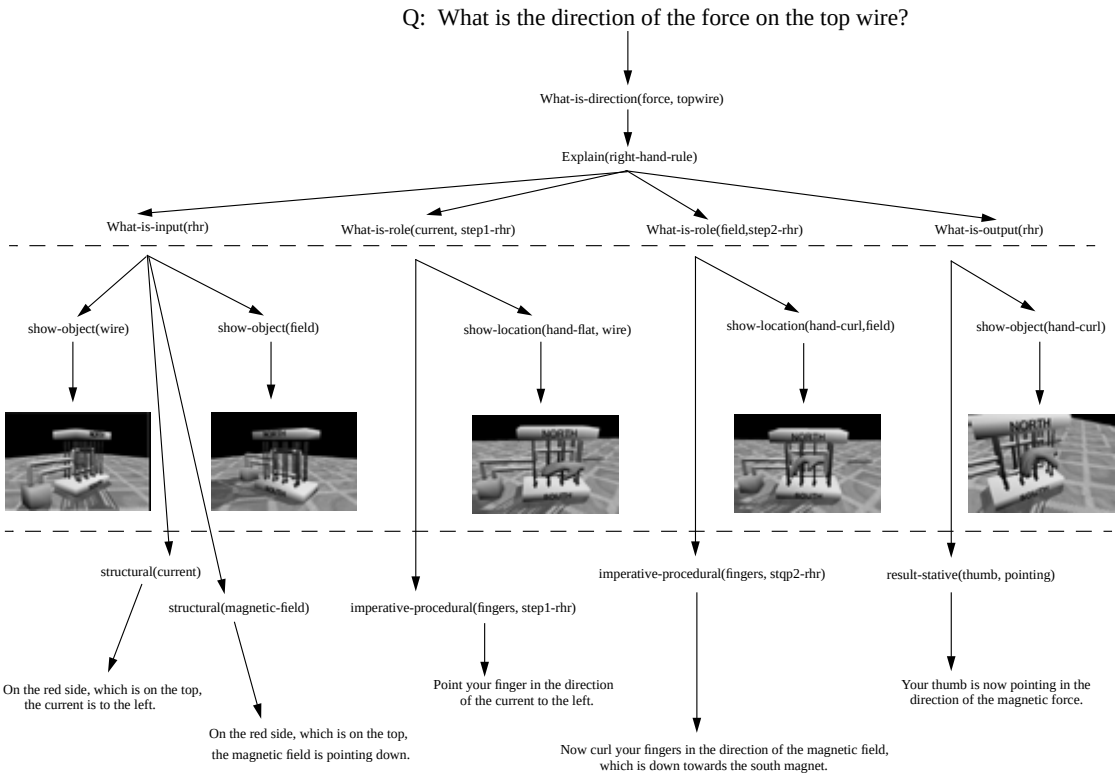

Figure 3: Example 3D multimedia explanation plan

animated explanations is a synthetic process of organizing the raw materials of 3D wire frame models and scene geometries and planning "camera shots" of the virtual camera:

1. *3D Model Selection:* Given a query which specifies a question type, e.g., `(explain-function ?X)`, and a target concept, e.g. `battery`, the explanation system uses the ontological indices of the knowledge base to retrieve the relevant *concept suite*. Indicating the most relevant visual and auditory elements, a concept suite is defined by a sequence of concepts, each of which is either an object, e.g., `Electrode` or a process, e.g., `Current-Flow`. The animation planner then selects the relevant wireframe models and introduces them into the virtual scene.

2. *Camera Shot Planning:* Through judicious camera shot selection, explanations can direct students' attention to the most important aspects of a scene, even in complex scenes presenting a number of objects in motion, and provide visual context. While high and far shots present more information (Mascelli 1965), close-up shots are useful for centering on a single subject (Millerson 1994). To provide visual context, it initially selects far shots for unfamiliar objects, unfamiliar processes, and tracking moving objects. It selects close-ups for presenting the details of familiar objects.

3. *Time Map Construction:* A time map houses parallel series of 3D coordinate specifications for all object positions and orientations, visual effects, and camera positions and orientations, with which the renderer can construct a frame of the explanation for every tick of the clock. These frames will be rendered with the accompanying narration in realtime, creating a continuous immersive visualization in which rich 3D explanations mesh seamlessly with the student's exploration of the environment.

Generating Spatial Natural Language Utterances

Given the spatial linguistic specifications created by the visuo-linguistic mediator, the natural language generator must utilize its grammar and lexicon to create sentences realizing the given content. The natural language generator copes with difficulties of producing spatial text by exploiting knowledge about position, direction, and orientation. It avoids utterances that otherwise would be spatially ambiguous by distinguishing the basic categories of spatial relationships that bear on objects in a three dimensional world. For example, the physics testbed for electromagnetism requires the language generator to ontologically discern the following in order to avoid spatial ambiguity:

- *Positions*: `left-side`, `top-side`, `bottom-side`, `right-side`, `center`.

- *Orientations*: `facing-up`, `facing-down`, `facing-left`, `facing-right`, `facing-toward`, `facing-away-from`.
- *Relative Orientations*: `perpendicular`, `parallel`, `oblique`.
- *Rotations*: `clockwise`, `counterclockwise`.
- *Curl Directions*: `curl-towards`, `curl-away-from`, `curl-up`, `curl-down`, `curl-left`, `curl-right`.

This family of spatial primitives enables the generator to appropriately adjudicate between a broad range of ambiguous candidate realizations. For example, although the position `left-side` and the orientation `facing-left` will be realized with the same lexicalization ("left"), the former case will occupy part of a noun phrase and the latter will be adverbial. With the linguistic specifications in hand, the natural language generator's sentence planner exploits the spatial ontology to map the given ontological concepts (e.g., `facing-left`) to the appropriate semantic role necessary to correctly realize the linguistic specification. For example, specifications frequently include features for relative position and pointing direction. These serve as cues to the natural language generator that enable it to distinguish the appropriate semantic roles. To illustrate, Figure 4 shows the result of a specification mapped to a *functional description* (Elhadad 1992). In this specification, the concept of left-side is realized as a locative semantic role because it had been marked in the specification as being a position type. If the primary actor had instead been a direction rather than a position, it would have been mapped not to a locative role but rather to a predicate-modifying adverb.

After the sentence planner constructs functional descriptions, it passes them to a unification-based surface generator (Elhadad 1992) to yield the surface string, which is itself passed to a speech synthesizer and delivered in synchronization with the actions of the associated 3D visualization. This process is repeated for each leaf of the explanation plan as the planner walks across the specifications in a left-to-right order. When the final verbal and visual elements of the explanation have been constructed, they are presented to the student, and the planner awaits the student's next question.

An Implemented Multimedia Explanation Generator

All of the components of the spatial explanation framework have been implemented in a realtime explanation planner that constructs integrated 3D animations and speech for complex three dimensional spatial phenomena. Given queries about directions, orientations, and spatial roles of forces, it generates 3D visualizations, produces coordinated natural language utterances, and synchronizes the two. The explanation planner is implemented in a heterogeneous computing environment consisting of two PentiumPro 200s and a Sparc Ultra communicating via TCP/IP socket protocols over

```
((CAT CLAUSE)
 (CIRCUM
   ((LOCATION
      ((POSITION FRONT)
       (CAT PP)
       (PREP ((LEX ''on'')))
       (NP ((CAT COMMON)
            (DEFINITE YES)
            (LEX ''side'')
            (DESCRIBER ((CAT ADJ)
                        (LEX ''red'')))
            (QUALIFIER
              ((CAT CLAUSE)
               (RESTRICTIVE NO)
               (SCOPE {^ PARTIC LOCATED})
               (PROC ((TYPE LOCATIVE)))
               (PARTIC ((LOCATION
                          ((CAT PP)
                           (PREP === ''on'')
                           (NP ((CAT COMMON)
                                (COUNTABLE NO)
                                (LEX ''top'')))))))
               (MOOD SIMPLE-RELATIVE)))))))))
 (PROC ((TYPE LOCATIVE)))
 (PARTIC
   ((LOCATED ((CAT COMMON)
              (DEFINITE YES)
              (LEX ''current'')))
    (LOCATION ((CAT PP)
               (PREP === ''to'')
               (NP ((CAT COMMON)
                    (DEFINITE YES)
                    (LEX ''left side'')))))))))
```

Figure 4: Example function description: current

an Ethernet. Both the media-independent explanation planner and mediator were implemented in the CLIPS production system. The 3D animation planner was implemented in C++. The spatial natural language generator was implemented in Harlequin Lispworks and employs the FUF surface generator and SURGE (Elhadad 1992), a comprehensive unification-based English grammar. The animation renderer was created with the OpenGL rendering library, and the speech synthesis module employs the Truetalk synthesizer. With regard to efficiency issues, the media-independent explanation planner, mediator, and animation planner operate in a small number of milliseconds; the natural language generator requires approximately 2–8 seconds, with the bulk of the time consumed by unification.

The PHYSVIZ Testbed

To study CINESPEAK's explanation planning behaviors, it has been incorporated into PHYSVIZ, a prototype 3D learning environment for the domain of high school physics. Physics presents a particularly challenging set of communicative requirements because many fundamental physics concepts are exceptionally hard to visualize. For typical physics students, attempting to understand these concepts by studying static two-dimensional graphics typically yields a less-than-satisfying learning experience. Focusing on concepts of electromagnetism, PHYSVIZ exploits a library of 3D models representing a battery, wires, magnets, and magnetic fields. It also includes a virtual 3D hand that can be used to explain the right-hand rule for determining the direction of magnetic forces. PHYSVIZ was developed by a multidisciplinary design team that in-

cluded an experienced high school physics teacher. In a numerous sessions spanning a semester, the physicist was posed detailed questions about electromagnetism. His verbal responses and his diagrammatic reasoning guided the design of the 3D models in the learning environment.

Example Explanation Planning Episode

To illustrate CineSpeak's behavior, suppose a student interacting with PhysViz constructs the query, "What is the direction of the force on the top of the wire?" The media independent explanation planner determines that it should create an explanation of the right-hand rule to respond to the question. There are four major steps in explaining the right-hand rule, which will be explained sequentially. It first explains the inputs, the current and the magnetic field and eventually proceeds on to the outcome of the right-hand's rule application, which is that the direction of the magnetic force is equivalent to the resulting orientation of the thumb. This content and the sequential organization are housed in the leaves of the media-independent explanation plan.

The mediator now coordinates the planning of animation and speech. First, the animation planner creates a 3D visualization plan consisting of specifications for the relevant 3D models (the wire, the magnetic field, and the virtual hand), their orientations, and relevant camera views that clearly depict these objects. Next, the mediator creates specifications for the natural language generator, continuing until an impasse is reached resulting from a dearth of up-to-date spatial information. It notes that the relative orientation of the current's direction is from right to left for this particular camera view. It requests and receives this information from the animation planner. It continues in this fashion until complete linguistic specifications have been created. It then passes the full specifications to the natural language generator, which creates a functional description for each.

Finally, the animation plan is passed to the renderer while the text string is passed to the speech synthesizer. As the renderer constructs a 3D visualization depicting the virtual hand pointing in the direction of the current (which it determines is to the left of the screen based on the student's vantage point), the speech synthesizer says, "Point your fingers in the direction of the current to the left." After explaining how the hand curls in the direction of the magnetic field, it concludes by visually demonstrating how the virtual hand's direction and orientation are used to determine the direction of the magnetic force on the top section of the wire while it says, "Your thumb is now pointing in the direction of the magnetic force."

Focus Group Study

To investigate the effectiveness with which CineSpeak generates clear 3D explanations of spatial phenomena, in addition to replicating the physicist's communication techniques (albeit in 3D but with more limited natural language phrasing), we conducted an informal focus group study with nine college-age subjects drawn from both technical and non-technical backgrounds. They were introduced to the PhysViz learning environment interface and briefly grounded in the basic concepts. Because many people unfamiliar with computer-generated speech frequently find it difficult to understand, subjects were first exposed to sample utterances produced by the speech synthesizer. Bearing in mind the caveat that the study was quite informal, results were nevertheless very encouraging:

- *Viewing perspectives:* Subjects unanimously liked the viewing perspectives chosen in the course of explanations and the dynamic highlighting of objects being referred to in the speech.
- *Timing and Synthesis:* Undoubtedly the most problematic aspect of the explanation stemmed from implementation limitations. Most subjects found the long delays between utterances, which were caused primarily by the time spent by the surface generator on unification, to be bothersome and the quality of the speech to be much less than ideal.
- *Superiority of Coordinated Multiple Media:* Perhaps the most telling finding was that the more redundancy between visual cues and verbal utterances, the more subjects understood the concepts. For example, explanations of current do not include visualizations of it other than the mere presence of the wire; explanations of current and its orientation were generated solely with verbal phrasings and an occasional use of the virtual hand. In contrast, explanations of magnetic fields, which employed both visual representations in the form of 3D arrows and magnets as well as verbalizations of the field orientation, were much more easily understood. Because subjects, unprompted, eagerly voiced their strong preferences for the latter over the former, the differences were particularly striking. This finding is consistent with a growing body of empirical evidence on the effectiveness of multiple modalities in intelligent multimedia interfaces, e.g., (Oviatt 1997).

Conclusion

The visuo-linguistic explanation generation framework can be used to create 3D multimedia explanations of complex spatial phenomena. By exploiting a mediator that serves as an intermediary between a 3D animation planner utilizing geometric spatial knowledge and a natural language generator that utilizes linguistic spatial knowledge, the visuo-linguistic explanation framework takes advantage of the strengths of both types of representations to generate clear spatial explanation combining 3D animations and complementary speech. In combination, well-designed visualizations integrated with spatial utterances effectively communicate complex three-dimensional phenomena. While this work provides a strong computational foundation

for generating integrated 3D animations and natural language, much remains to be done, particularly with regard to generating 3D spatial explanations of highly dynamic phenomena. This entails extending the animation planner's ability to render 3D models exhibiting more complex behaviors, the natural language generator's dynamic spatial linguistic coverage, and the visuolinguistic mediator's arbitration strategies for coordinating more dynamic spatial knowledge. We will be pursuing these activities in our future work.

Acknowledgements

The authors gratefully acknowledge Dr. Loren Winters of the North Carolina School of Science and Mathematics for collaboration on all aspects of the PhysViz learning environment, William Bares for his generous assistance in integrating the Rapid 3D animation and cinematography system, and Luke Zettlemoyer for his work in preparing the manuscript.

References

André, E., and Rist, T. 1996. Coping with temporal constraints in multimedia presentation planning. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 142–147.

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.

Burton, R. R., and Brown, J. S. 1982. An investigation of computer coaching for informal learning activities. In Sleeman, D., and Brown, J. S., eds., *Intelligent Tutoring Systems*. London: Academic Press. 79–98.

Butz, A., and Krüger, A. 1996. Lean modeling—the intelligent use of geometrical abstraction in 3D animations. In *Proceedings of the Twelfth European Conference on Artificial Intelligence*, 246–250.

Cawsey, A. 1992. *Explanation and Interaction: The Computer Generation of Explanatory Dialogues*. MIT Press.

Christianson, D. B.; Anderson, S. E.; He, L.-W.; Salesin, D. H.; Weld, D. S.; and Cohen, M. F. 1996. Declarative camera control for automatic cinematography. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 148–155.

Davis, E. 1990. *Representations of Commonsense Knowledge*. San Mateo, CA: Morgan Kaufmann.

Elhadad, M. 1992. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation*. Ph.D. Dissertation, Columbia University.

Feiner, S. K., and McKeown, K. R. 1993. Automating the generation of coordinated multimedia explanations. In Maybury, M. T., ed., *Intelligent Multimedia Interfaces*. Menlo Park, CA: AAAI Press/The MIT Press. chapter 5, 117–138.

Feiner, S. 1985. APEX: An experiment in the automated creation of pictorial explanations. *IEEE Computer Graphics and Applications* 29–37.

Fillmore, C. 1975. *Santa Cruz Lectures on Deixis 1971*. Available from Indiana University Linguistics Club.

Gapp, K.-P. 1994. Basic meanings of spatial relations: Computation and evaluation in 3D space. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 1393–1398.

Hollan, J. D.; Hutchins, E. L.; and Weitzman, L. M. 1987. STEAMER: An interactive, inspectable, simulation-based training system. In Kearsley, G., ed., *Artificial Intelligence and Instruction: Applications and Methods*. Reading, MA: Addison-Wesley. 113–134.

Hovy, E. H. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence* 63:341–385.

Karp, P., and Feiner, S. 1993. Automated presentation planning of animation using task decomposition with heuristic reasoning. In *Proceedings of Graphics Interface '93*, 118–127.

Lesgold, A.; Lajoie, S.; Bunzo, M.; and Eggan, G. 1992. SHERLOCK: A coached practice environment for an electronics trouble-shooting job. In Larkin, J. H., and Chabay, R. W., eds., *Computer-Assisted Instruction and Intelligent Tutoring Systems: Shared Goals and Complementary Approaches*. Hillsdale, NJ: Lawrence Erlbaum. 201–238.

Mascelli, J. 1965. *The Five C's of Cinematography*. Cine/Grafic Publications, Hollywood.

Maybury, M. T. 1994. Planning multimedia explanations using communicative acts. In *Proceeding of AAAI-91*, 65–66.

Millerson, G. 1994. *Video Camera Techniques*. Focal Press, Oxford, England.

Mittal, V.; Roth, S.; Moore, J. D.; Mattis, J.; and Carenini, G. 1995. Generating explanatory captions for information graphics. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1276–1283.

Moore, J. D. 1995. *Participating in Explanatory Dialogues*. MIT Press.

Novak, H.-J. 1987. Strategies for generating coherent descriptions of object movements in street scenes. In Kempen, G., ed., *Natural Language Generation*. Dordrecht, The Netherlands: Martinus Nijhoff. 117–132.

Oviatt, S. 1997. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction* 12:93–129.

Rieber, L. 1990. Animation in computer-based instruction. *Educational Technology Research and Development* 38(1):77–86.

Roth, S. F.; Mattis, J.; and Mesnard, X. 1991. Graphics and natural language as components of automatic explanation. In Sullivan, J. W., and Tyler, S. W., eds., *Intelligent User Interfaces*. New York: Addison-Wesley. 207–239.

Seligmann, D. D., and Feiner, S. K. 1991. Automated generation of intent-based 3D illustrations. *Computer Graphics* 25(4):123–132.

Sibun, P. 1992. Generating text without trees. *Computational Intelligence* 8(1):102–122.

Suthers, D. D. 1991. A task-appropriate hybrid architecture for explanation. *Computational Intelligence* 7(4):315–333.

Wahlster, W.; André, E.; Finkler, W.; Profitlich, H.-J.; and Rist, T. 1993. Plan-based integration of natural language and graphics generation. *Artificial Intelligence* 63:387–427.