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

Charles B. Callaway and **Stuart G. Towns** and **James C. Lester**

Multimedia Laboratory
Department of Computer Science
North Carolina State University
Raleigh, NC 27695

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. 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 multimedia explanations of three dimensional electromagnetic fields, forces, and electrical current in realtime.

Introduction

As multimedia technologies reach ever higher levels of sophistication, knowledge-based learning environments and intelligent training systems can create increasingly effective educational experiences. 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 (Neal & Shapiro 1991; Maybury 1994; 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 & 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 can serve as a mediator. This framework has been implemented in CINESPEAK, a multimedia explanation generator consisting of a media-independent explanation planner, a visuo-

* Extended versions of this paper have been submitted to INLG-98 and AAAI-98. Support for this work was provided by the National Science Foundation under grant IRI-9701503 (CAREER Award Program), the William R. Kenan Institute for Engineering, Technology and Science, the North Carolina State University IntelliMedia Initiative, and Novell.

linguistic mediator, a 3D animation planner, and a real-time natural language generator with a speech synthesizer. CINESPEAK has been used in conjunction with PHYSVIZ, a prototype 3D learning environment in the domain of physics, to generate multimedia explanations of three dimensional electromagnetic fields, forces, and electrical current in realtime.

Spatial Explanation Generation

A critical functionality of knowledge-based learning environments and training systems is automatically providing students with clear explanations of spatial phenomena. Generating clear spatial explanations entails addressing three 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, 3D animations without accompanying explanatory speech are too limiting. While previous work has addressed the coordination of 2D graphics and natural language (Neal & Shapiro 1991; 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.

- *Dual Representation of Geometric and Linguistic Spatial Knowledge:* While we are far from a comprehensive theory of spatial reasoning, which must include 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.

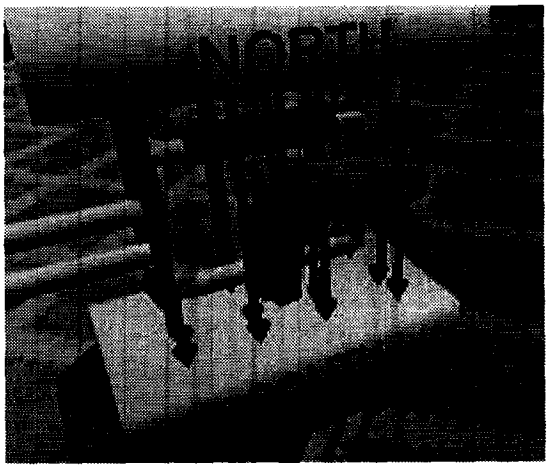

Figure 1: Explaining electromagnetism in the PHYSVIZ learning environment

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 takes the goal and constructs 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. The animation planner 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. The language generator 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 greatest 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-

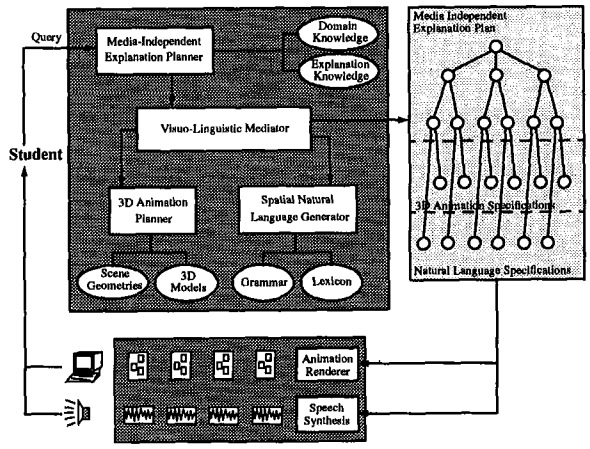

Figure 2: The visuo-linguistic explanation framework

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 a somewhat simplified version of the by-now classical top-down decomposition approach to explanation generation (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. For example, the content in the example above is organized in the structure depicted in the second level of the explanation plan 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 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.
- *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 are 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) It 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 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

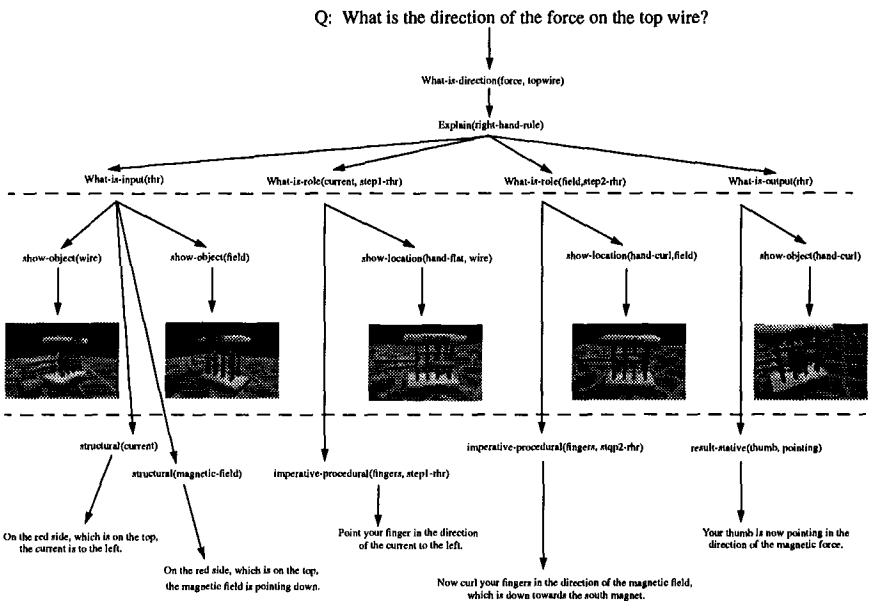

Figure 3: Example 3D multimedia explanation plan

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. Planning 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]

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. 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.

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

Generating Spatial 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`.

```
((CAT CLAUSE)
 (CIRCUM
  ((LOCATION
    ((CAT PP)
     (PREP ((LEX ''on'')))
           (POSITION FRONT)
           (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 spatial functional description

- *Orientations*: `facing-up`, `facing-down`, `facing-left`, `facing-right`, `facing-toward`, `facing-away`.
- *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. Figure 4 shows the result of a specification mapped to a *functional description*. 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.

An Implemented Multimedia Explanation Generator

All of the components of the spatial explanation framework have been implemented in CINESPEAK, a realtime explanation planner that constructs integrated 3D animations and speech for complex three dimensional spatial phenomena.[2] Given queries about directions, orientations, and spatial roles of forces, it generates 3D visualizations, produces coordinated natural language utterances, and synchronizes the two.

The PHYSVIZ Learning Environment

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. Focusing on concepts of electromagnetism, PHYSVIZ exploits a library of 3D models representing a battery, a wire, magnets, and a magnetic field. It also includes a virtual 3D hand that can be used to explain the right-hand rule for determining the direction of magnetic forces.

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 rule's 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 the 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

[2]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 an Ethernet. Both the media-independent explanation planner and mediator were implemented in a 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). The animation renderer was created with the OpenGL rendering library, and the speech synthesis module employs the Truetalk synthesizer.

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 spec. 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), an informal focus group study was conducted. Nine college-age subjects were drawn from both technical and non-technical backgrounds. 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 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 so 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.

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.

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.

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.

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

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.

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

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

Neal, J. G., and Shapiro, S. C. 1991. Intelligent multimedia interface technology. In Sullivan, J. W., and Tyler, S. W., eds., *Intelligent User Interfaces.* New York: Addison-Wesley. 11–43.

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.

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

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.