Language-Based Interfaces and Their Application for Cultural Tourism

Oliviero Stock

■ Language processing has a large practical potential in intelligent interfaces if we take into account multiple modalities of communication. Multimodality refers to the perception of different coordinated media used in delivering a message as well as the combination of various attitudes in relation to communication. In particular, the integration of natural language processing and hypermedia allows each modality to overcome the constraints of the other, resulting in a novel class of integrated environments for complex exploration and information access. Information presentation is a key element of such environments; generation techniques can contribute to their quality by producing texts ex novo or flexibly adapting existing material to the current situation. A great opportunity arises for intelligent interfaces and language technology of this kind to play an important role for individual-oriented cultural tourism. In the article, reference is made to some prototypes developed at IRST that were conceived for this specific area. A recent project concentrated on the combination of two forms of navigation taking place at the same time—one in information space, the other in physical space. Collaboration, an important topic for intelligent interfaces, is also discussed.

Language is the extraordinary means provided by the human mind for communicating with other humans (and for structuring thoughts). For a long time, spoken language has been the means for communicating face to face. Written language, a means for transporting language across space and time, was invented about 5000 years ago, most likely by the Sumers. In the beginning, written language was pictorial, and after a few centuries, cuneiform coding was proposed. Only much later, in the thirteenth century BC according to some archaeologists, was an alphabet introduced (in Ugarit, in what now is Syria). It consisted of about 30 cuneiform signs and was quickly recognized as a breakthrough and adopted by several peoples.

Means for producing various instances of a written "document" were invented soon after the first appearance of written language, but it took more than 4000 years for Gutenberg to create his flexible printing system based on characters of the alphabet. It took 500 more years to get to the computer and its possibilities. Shortly before that, some other means for long distance communication (for example, the telegraph or the telephone) also appeared and, in being adopted, produced some slight variation between the modalities of the spoken and the written word.

With the computer, the flexibility in dealing with the form of written language (editing) and accessing language (retrieving) is emphasized. However, in the scientific area of natural language processing, the goal is much more ambitious: to automatically understand and produce language. Being potentially able to deal with the content of the message has opened the way to communicating with a machine through language. In particular, effort has been put into making it possible to interact with a computer to get some desired information. Although the results have been fairly significant, their impact has been scarce when we consider how society might benefit if computers could understand human language, making information more accessible. It took quite a lot for the scientific community to understand that communicating with a com-

Figure 1. A Representation of Ambiguity in Multimodal Communication (The Calling of St. Matthew by Caravaggio, Rome, 1600).

puter through natural language might mean something different from the two basic language modalities (Maybury and Wahlster 1998). The so-called *teletype approach* (a very limited view of the interface with a substantial narrow bandwidth of communication) has persisted for some time. Later, we began to understand that a larger bandwidth of communication can be established between human and computer. For example, language can be integrated with images dynamically; the screen itself is not only an output medium but can become the basis for direct manipulation of all objects involved in the communication (through a pointing device or a gesture-recognition device) (Maybury 1993).

However, the point is not only in the interface. Often the user does not know what information is available to him/her, or he/she might not have a clear idea of what he/she is searching for. The need arises for systems that integrate a mediated information access paradigm and a navigational paradigm, where the user might use different modalities to explore the material in a way possible only with a computer. We believe that exploration of an information space will become an increasingly typical interactive attitude for users, in a world inhabited by a multitude of available multimedia information (Maybury 1997). Such exploration is becoming apparent with the current diffusion of the web and its various browsers.

Another key element of flexibility lies in the possibility of a system having a model of the user, including his/her interests, idiosyncrasies, and the dynamic aspects inferred during the interaction. This model is instrumental in making sense of partial or insufficiently detailed requests (or other acts) by the user and for determining the system's actions.

A desirable feature is the appropriate presentation of information; that is, the relevant information is made available to the user at the proper level of detail, coherent with other

pieces of information provided previously, and further exploration is favored.

If information is to be presented in a flexible way, it is essential that an automatic processor do the job—in the case of text presentation, a natural language–generation system. Given the internal representation of the knowledge sources, the system decides what the relevant information is to be communicated, organizes a coherent text structure, and produces the most appropriate linguistic expressions to convey the message. Multimodal flexible presentations exploit synergistically the advantages that different media can provide in conveying the message to the user. In this case, all processors must start from an internal representation, and the system must organize media allocation and media coordination.

I discuss in turn some of these themes, making reference to work we have developed at the Institute for Scientific and Technological Research (IRST) in Trento, Italy.

Multimodality and Exploration of Information

There can be different views on multimodal communication. Multimodality as such is multidimensional. Often it is regarded only as the combination of various uses of media, but certainly this combination is only one obvious aspect of the whole matter, one that requires a clarification. *Multimedia* denotes the physical means by which information is input, output, and stored. *Multimodality* refers to the human perceptual processes such as vision, audition, and taction and can also refer to the interpersonal or person-artifact context that develops in the interaction. Intelligent (multimodal) systems in principle tend to be characterized by a representation of the content of the presentation, so that presentation material is not fixed and can be customized dynamically.

In a masterpiece painted in 1600 and visible in the church of San Luigi dei Francesi in Rome (figure 1), Caravaggio, the great painter, has represented genially some of the problems in multimodality and, in particular, in gesture understanding. In the scene, Jesus is calling someone seated in an inn and points to him. However, the pointing act is ambiguous. Saint Peter, standing next to him, reinforces the indication by Jesus, but his pointing act is uncertain and does not help. The person seated to the left recognizes the pointing act and seems to ask, "Do you mean me?" However, the viewer might think he is indicating the person next to him laying over the table half asleep or drunk. Caravaggio got it right. The confusion is total if language does not have a key role and if there is no interactivity and feedback.

Multimodality can also refer to the combination of various attitudes in relation to communication and information access (for example, goal oriented and exploration oriented), each having its own specific characteristics. Another view of multimodality relates to human capabilities. It can amplify our capabilities. To do so, it must be cognitively compatible, but altogether, it might not reflect our "natural" communication. Yet another aspect of multimodality is related to the specific role that language plays in the context (Stock 1995). It is a different role than when communication is based on a single modality, so its operational characteristics are different; its modeling requires specific components.

Integration of natural language processing and hypermedia in a multimodal system offers a high level of interactivity and system habitability; each modality overcomes the constraints of the other, resulting in a novel class of integrated environments for complex exploration and information access.

According to Waterworth and Chignell (1991), there are at least two dimensions for a model of information exploration: (1) structural responsibility and (2) target orientation. *Structural responsibility* involves the issue of which agent (that is, the user or the system) is responsible for carrying out search and giving structure to information. It gives rise to a dichotomy between navigational and mediated exploration. The dimension of target orientation presents a dichotomy between browsing and querying. *Browsing* is distinguished from querying by the absence of a definite target in the mind of the user. This distinction is determined only by the cognitive state of the user, not by his/her actions or the configuration of the system. In reality, there is a continuum of user behaviors varying between querying and browsing, so it is inappropriate to build systems that reflect this strict dichotomy, imposing one particular attitude on the user's exploration.

In work carried on for several years at IRST, we developed an environment in which interaction could smoothly move along the two dimensions. Dialogue management had to include a communicative action coordinator, responsible for using media properly (so, for example, it can take into account the deictic context at any time of the interaction) and/or suggesting to the user a shift along the structural responsibility dimension.

These ideas are at the basis of the ALFRESCO interactive system (Stock 1991). For this system

Multimodality can also refer to the combination of various attitudes in relation to communication and information access (for example, goal oriented and exploration oriented), each having its own specific characteristics.

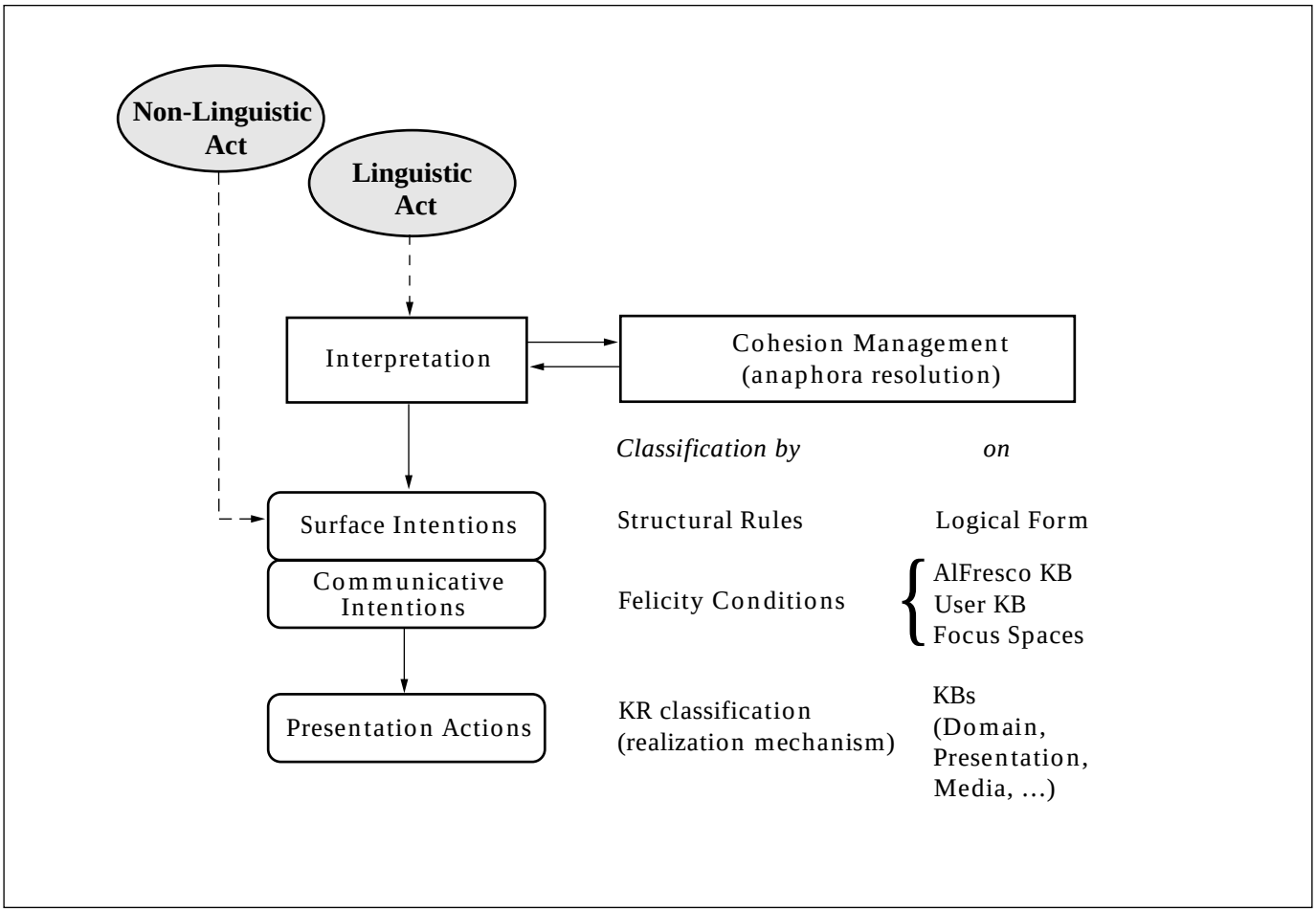


Figure 2. ALFRESCO Dialogue Management.

and for other research I report on here, there is a common application theme: cultural heritage and tourism. This is no surprise when you consider that Italy is believed to have half of the world's cultural tourism resources. The field can provide a wonderful opportunity for introducing technology that can help the shift from a mass-oriented attitude to an individual-oriented attitude—exactly what we aim for. Cultural tourism can become an experience in which the individual is the active subject of the exploration, one who develops a personal taste and interest.

ALFRESCO is an interactive, natural language–centered system for a user interested in fourteenth-century Italian frescoes. It has the aim of providing information and promoting other masterpieces that might attract the user. Hypermedia is integrated both in input and output. The user can interact with the system by typing sentences, navigating in an underlying hypertext, and using the touch screen in a coherent multimodal discourse setting. In output, images and generated text offer entry points for further hypertextual exploration. The result is that the user communicates linguistically and manipulates various entities, images, and text. The system builds a simple model of the user as the dialogue proceeds and uses it for output decisions, still allowing the user to browse around freely.

A higher-level, pragmatic component decides how to react in the given dialogic situation, considering the type of utterance by the user, the context, the model of the user's interest, the things already shown or said to the user, and so on. The dialogue can result in zooming onto details or changing the focus of attention to other frescoes.

We have proposed a level of multimodal act representation (Stock, Strapparava, and Zancanaro 1997), roughly corresponding to, for strictly linguistic dialogues, the illocutionary level (figure 2). The key point for multimodal interaction is provided by the uniform use of *felicity conditions*, the rules that govern the relations between interactional exchanges and communicative intentions.

Felicity conditions use a model of the user's knowledge (what the system guesses the user knowledge is) and the changes of focus spaces in the multimodal environment. For example, they help in determining if the user is making a request to (1) describe a particular entity (in this case, the user already knows which entity satisfies the constraints but has not received further information), (2) find some information about a particular entity (he/she must not know which entity satisfies the constraints), and (3) locate an entity among a group of entities (in the focus of attention, there must be an entity satisfying the constraints, and the user must know so). Then a question such as "Who is the author?" would produce different output in different situations (for example, if a fresco was shown before, and its contents have now been described without mentioning the name of the author or if the fresco has just been shown, and the author had represented himself among other characters). I remember an initial demonstration of ALFRESCO before this component was added. An important visitor asked "Who is Giotto?" After processing for a long time, the system came out with the reply: "Giotto." This statement was not exactly helpful.

The dialogue cohesion management (Zancanaro, Stock, and Strapparava 1997b, 1993) also provides the user with a graphic feedback of the dialogue cohesion status. This visual representation (1) reassures the user at a glance about the system's interpretation (as such, it takes the place of a paraphrase) and (2) allows cooperative recovery from discourse misconceptions by means of a series of "intuitive actions" when this interpretation is not the one the user meant. In general, we have tightly integrated different modes of exploration: the language oriented and the navigational. I think this concept is very fruitful and can lead to various applications. The study of the involved cognitive aspects is of great importance, and lab experiments with implemented prototypes and simulated systems will make all of us better understand this kind of amplification of human communicative capabilities.

Bringing Physical Space into the Picture

One further element for advancing in the direction of personalization and context sensitivity is offered by ubiquitous information access, made possible by hardware technologies such as portable devices and wireless networking. A museum is a privileged environment for introducing adaptive information with ubiquitous access. In fact, the experience of visiting a museum typically consists of moving in a physical space and acquiring information about the objects shown (and, of course, becoming interested and moved by what is displayed!). In the new interaction scenario, the computer (a hand-held device including spoken output) allows the integration between the physical space (through a positioning system) and the related information space, yielding a new way of exploring cultural heritage. The individual visitor is at the center of the physical-virtual space exploration, and his/her movements and interactions provide input to the system to tailor appropriate presentations.

The approach presented here was developed inside a project at IRST called HYPERAUDIO (Not et al. 1998). The results are at the basis of the development of an even richer interaction scenario that is being explored jointly with other partners in HIPS (HYPERINTERACTION WITHIN THE PHYSICAL SPACE), a European project of the Esprit I^3 (intelligent information interfaces) program (Benelli et al. 1999).[2] The problem of adapting content for (cultural) information presentations in physical hypernavigation shares many features with the problem of producing adaptive and dynamic hypermedia for virtual museums (for example, ILEX [Mellish et al. 1997]) or dynamic encyclopedias (for example, PEBA-II [Milosavljevic, Tulloch, and Dale 1996]). Moving in a physical museum has been the goal of the RHINO project, where a robot accompanies the visitor (Burgard et al. 1998). A fascinating work on wearable augmented reality systems that include localization, vision, and graphics and caption overlay for a person moving in a cultural outdoor environment is described in Höllerer, Feiner, and Pavlik (1999).

A museum is a privileged environment for introducing adaptive information with ubiquitous access.

Content adaptation in a physical environment poses some problems that are related to the visitor experiencing a "real" situation: The cognitive problems that might arise when a person is moving in a virtual information space are different when the user is seeking concrete objects and moving in a real environment that provides stimuli, attention grasping, and feedback. Information is presented in different situational contexts, determined mainly by (1) what the user position and movements are, (2) what the structure of the surrounding physical space (for example, whether objects are close) is, (3) whether other people are examining the same item, and (4) whether the user is alone.

HYPERAUDIO (and HIPS) integrates the individual, dynamic modeling of the user with a general model of the environment, the user's movements, and the discourse history to best tailor information presentations. Different

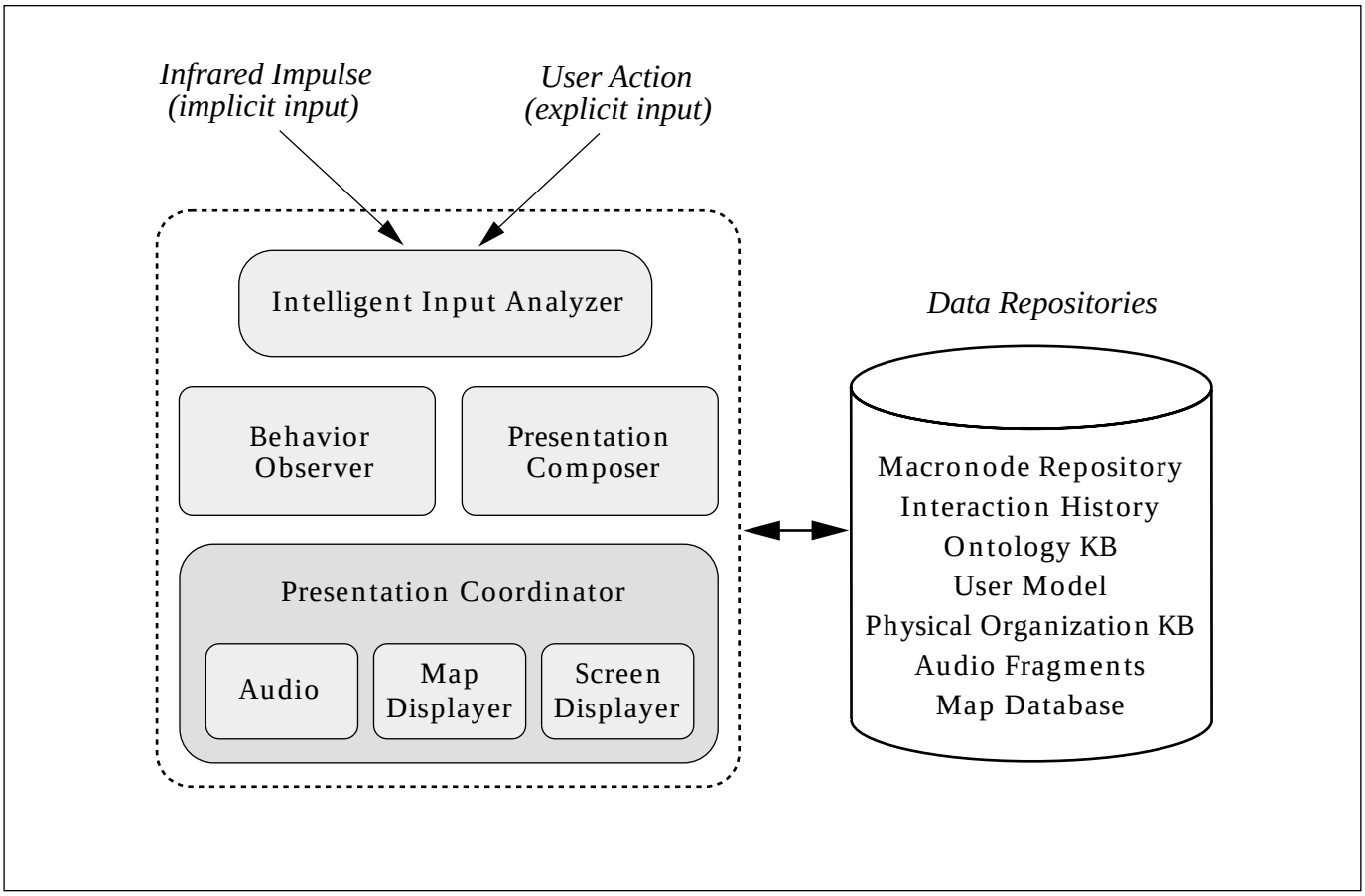


Figure 3. Architecture.

forms of adaptation are introduced by the system, both in the information provided and in the further steps suggested. In general, the approach points to a realistic and evolutionary adoption of generation techniques; at present, it yields a rhetorically coherent dynamic combination of small existing fragments of speech.

The architecture abstracts away from specific implementation solutions. It can be implemented on a single mobile platform (as in HYPERAUDIO) or with some modules running on a standing platform and communicating with the mobile computer by wireless connection (this solution is investigated within the HIPS project).

When deciding what information to include in the presentation and the most suitable discourse structure, the system takes into account various knowledge sources about the user and the interaction (figure 3).

The user model is accessed to exploit (1) the user's interests (which are inferred from his/her behavior) to include information in the new message that can stimulate the hearer's attention, possibly proposing information about other objects or sites strictly related to what the user is seeing, to increase curiosity and the desire to explore and (2) the user's background knowledge to relate the new information presented to what he/she already knows (therefore reinforcing learning) and decide whether additional clarification or exemplification of new concepts is required to help him/her understand. As the interaction proceeds, the system refines its assumptions about the user's interests and knowledge by observing the user's behavior and keeping track of the information to which he/she has been exposed. Another knowledge source is the history of previous interaction. An important role in content selection is also played by discourse strategies that the system exploits to guarantee that topics are presented in a coherent order, and the various discourse chunks are linked by rhetorical relations that reinforce the understanding of discourse flow. The system consistently limits the length of audio messages, deciding to realize part of the content as "clickable" links on the screen to avoid overwhelming the visitor with information. The language style adopted for

each user is selected according to his/her general visiting style. The user is dynamically classified along various dimensions on the basis of his/her behavior (Marti et al. 1999), and the rules in the PRESENTATION COMPOSER establish the appropriate amount of information to be presented.

According to information contained in the current context of interaction (for example, position in relation to displayed objects, the more extended environment, and so on) and in the discourse history (for example, the topic of the previous sentence or presentation), the system selects an appropriate linguistic realization for referring expressions and spatial references. Other cohesion devices (such as anaphora, conjunctions, lexical cohesion) are properly introduced to guarantee the fluency of the message and enhance understanding.

The system also includes a graphic interface that helps to orient the visitor and is useful for complementing linguistic instructions. Besides, "clickable" elements in the oral presentations appear on the screen. In figure 4, the enhanced version of the interface is shown. It is not yet on a specific hardware but is simulated on a portable PC and includes three-dimensional images.

Input provided by the user to the system can be both implicit, corresponding to movements in the physical space, and explicit, corresponding to interaction with the palmtop screen. Input is first analyzed by the intelligent input analyzer that decides on the most suitable type of processing required (for example, plan a new presentation, stop the current presentation, and plan a navigation support message). From all input, the behavior observer derives possible refinements to the user model.

The presentation composer is responsible for planning an overall presentation that integrates (where appropriate) object descriptions, images supporting descriptions, buttons and menus for follow-up information requests, directions for navigation support, and maps. The fundamental resource for flexible output generation is the *macronode repository*. Each macronode includes a network of message fragments (audio, text, or images), a list of pointers to other relevant macronodes (specifying the particular rhetorical relations among them), the type of message (for example, introductory label or caption), and a pointer to the relevant semantic concepts in the ontology. The network of message fragments encodes the different ways in which it is possible to realize the content of the macronode, thus encoding its surface linguistic forms, and the relevant concept, the rhetorical links, and the message type encode the deep structure of the description.

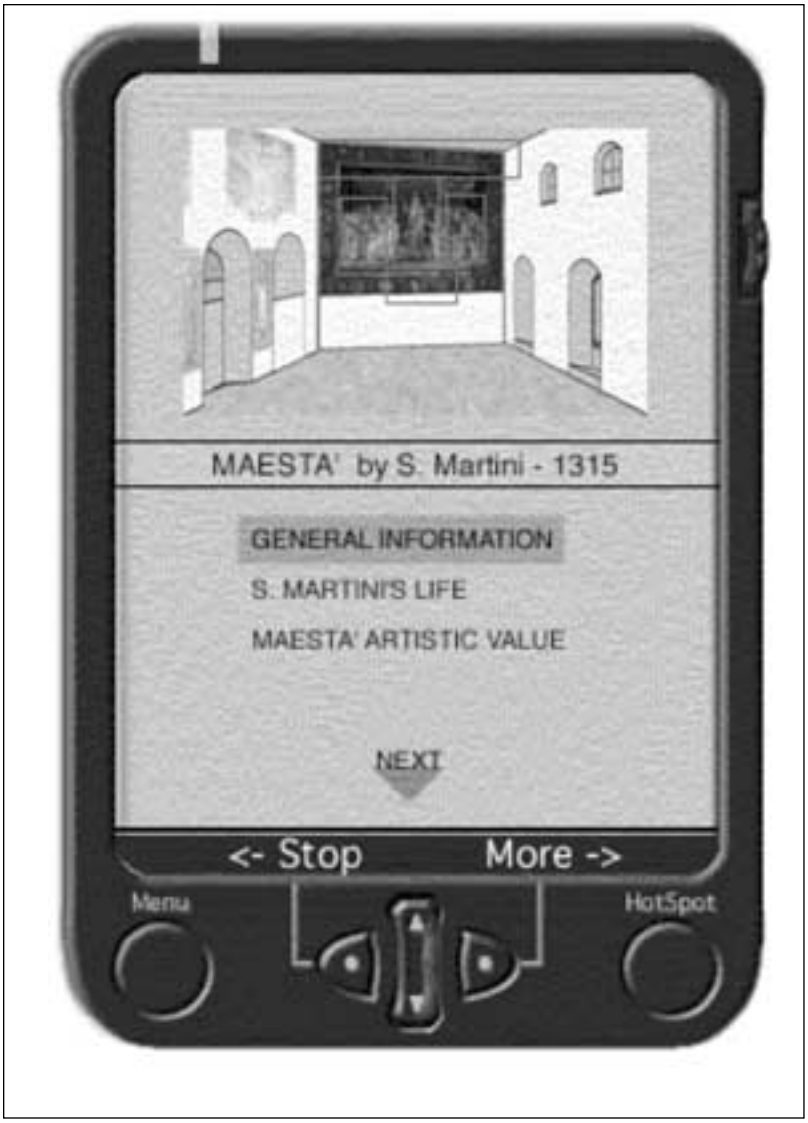


Figure 4. The Graphic Interface.

The main component of the presentation composer is a rule-based engine with three clusters of rules: The first cluster (*select start*) aims at the selection of the main macronode, given the object to talk about; the second cluster (*traversal_repository*) deals with how to traverse the macronode repository following the rhetorical relations to collect other macronodes and enrich the current presentation; the third cluster (*collect_follow-up*) specifies the rules for selecting hyperlinks for further, follow-up presentations. Each rule in a cluster is composed of a condition, written in a declarative Prolog-like language, and an output variable for controlling the result. Operators in the condition part of a rule can test all the knowledge sources (that is, the user model, the visit-

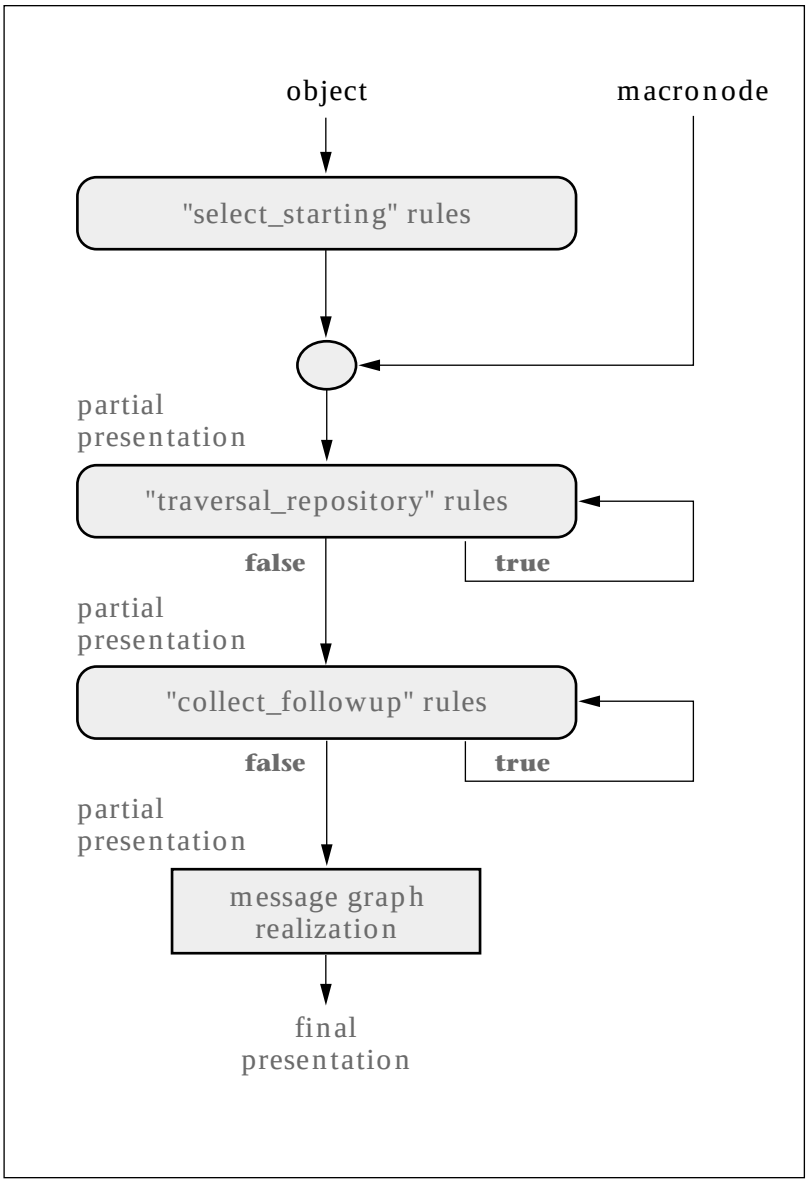


Figure 5. Presentation Composer.

ing style model, the physical organization knowledge base, the interaction history, and the macronode repository).

As shown in figure 5, clusters of rules are used iteratively to produce incrementally more detailed partial presentations. Finally for each macronode selected, a path in the message graph is chosen, checking the conditions against the discourse context and adjusting the presentation accordingly, for example, using connectives or referring expressions. This step of the presentation composition ensures the cohesion of the current presentation, and coherence is taken care of by the rest of the process.

Many of the issues presented for the museum setting apply to any physical hypernavigation setting in which individual, dynamic guides are appropriate, for example, historic cities; archaeological sites; or natural settings such as gardens, parks, or mountains. It is obvious that wide-open spaces introduce additional options from the technological point of view (for example the adoption of a global positioning system as the localization system) and suggest more ambitious scenarios (for example, new functions to support groups of visitors, access to online services such as weather forecasts). Another element being explored in HIPS is the adoption of global strategies for presenting information and promoting items, making sure the visitor does not miss them. The simplest strategy is a *gravity-driven* one that includes a basic path where the visitor, through presentations and suggestions, is attracted, despite any deviations he/she performs; distances from a position to the next position tend to be minimized. Another strategy brings into the picture the dimension of play, for example, a dynamic treasure hunt, where, typically, physical distances tend to be maximized.

Still another innovative feature is the introduction of collective memories. The visit trace is saved. The data can be accessed later by the visitor: When he/she is back home, he/she will be able to further explore the domain to which she had earlier been exposed with a system (such as ALFRESCO) that recalls the visit and will support him/her in successive exploration.

Other possibilities are there for treasuring some specific itinerary (for example, one made by an art critic or a public person) so that it can be followed with minor deviations by another visitor. Still another opportunity is to build models of the behavior of classes of visitors and, on this basis, influence the curators' choices.

What about Speech?

Speech processing is a key element for natural interaction systems. I believe that *synthesis* (the production of speech, normally starting from text) in particular will prove even more important than recognition. Often, the user's input can be simple (as in HIPS), but still output, because it can depend on other implicit input or on a user profile, can require a lot of sophisticated processing for achieving a good presentation level. With personalized output, you often want information to be presented as a coherent text, prepared for you, and presented orally. *Concept to speech* (deep integration of generation and synthesis) is yielding good results, but even synthesis as such has improved tremendously.

As for spoken input, there has been a lot of

Intelligent Interfaces for the Tourist

The overall phases of the individual tourist activity can be arranged in this way: (1) exposure to promotion, idea, and curiosity formation; (2) information access; (3) decisions and actions; (4) preparation for the visit; (5) the visit; and (6) further information exploration after the visit.

Current commercial technology is beginning to provide some access to online information and some systems for resource reservation and commercial transaction management for the tourist. Various tools can assist the user in negotiating and making the right choices so that certain tasks can be performed, such as making reservations or commercial transactions. The primary needs go much further, however; first is providing information of the appropriate quality in different languages and formats. However, the scarcest resource for our society in the new century will be people's time and attention, requiring a technology able to provide adequate instruments for interacting with the individual; inform, persuade, and so on; and offer new modalities of interaction. The most interesting topic is the individual experience, the visit and its consequences. Intelligent, language-based interfaces will contribute to the development of a new level of cultural tourism.

Scenarios

Hypermedia systems for tourism (or for museum exploration) are rather common at present. They exploit multimedia and virtual reality technologies but, in general, provide rather limited possibilities. The main weakness lies partly in the difficulty of making precise, complex requests in the information space and partly in the impossibility of pursuing specific aims through communication with the system. Interest for a destination and its culture grows slowly and autonomously, and just imposing more data or experiences than a person is prepared to cope with has no positive effect at all. It is essential to provide the user with the possibility to drive the game and explore culture according to his/her interest.

Before the visit: Exploring the information space—A cultural visit often begins before the departure for the chosen destination. The visitor might have read an article about an exhibition, read books and tourist guides about a town, seen videos, or obtained the advice of friends and experts. Nowadays, all these activities have a technological version: web-accessible digital libraries, video clips on demand, web newsgroups, and so on. However, the advantages offered by these technologies are often minimal, given the width of the information space. The user needs to adopt various integrated modalities—goal oriented and exploration oriented—for accessing information.

During the visit: Physical space and augmented reality—The experience of a visit to a museum or an art city typically consists of moving in a physical space and acquiring information about the objects one has met. Palmtop computers and wireless networks can help substantially innovate the way information is made accessible to the visitor—right at the moment it is necessary, presented through the most adequate channel (for example, audio or video) and in the form that best suits the user's interests and knowledge. The visitor, equipped with a wearable computer supplied with localization devices, moves freely in a museum or a city. The computer allows integration between real physical space and information space: The visitor is an active explorer, and his/her movements and interactions provide indications to the system for dynamically producing presentations in the specific situation. The system exploits and updates the model of the visitor's knowledge and interests initially developed at home.

After the visit: Taking home what one has seen—Personalized catalogues can automatically be built before and during the visit and completed at home, connected to the multimedia web sites of the museum or the town. The visitor's model built during the visit will allow him/her to continue to interact from home and further explore what he/she has been exposed to. The home system will know what he/she saw and what is interesting for him/her; it will keep her "connected" to new developments, continuing a "relationship" with the physical place he/she visited.

progress in spontaneous speech recognition, albeit in highly constrained dialogic settings. For example, within the C-STAR II consortium, IRST and its partners have built a prototype aimed at making it possible for two persons who are physically remote and each speaking his/her own language (Cettolo et al. 1999) to entertain a conversation oriented to book a hotel room. Again, the application is relevant for the tourism domain, and translation is the most apparent result, but probably the really important technological progress in input is in the ability to treat natural speech phenomena, such as false starts and hesitations.

Collaboration

What has been described so far is not enough. Dialogue must be seen as a collaborative enterprise, and we need models that help us understand multimodality at this deeper level. Our overall systems at IRST do not make use of this

Macronodes

The *macronode* formalism provides a way of annotating a set of existing repositories of data with the aim of making explicit the content and the relations that hold among different pieces of information.

For textual information, a macronode roughly corresponds to a paragraph. The annotation encompasses the description of the information contained in each unit, the relations with other units, and the different ways in which each unit can be presented to the user. Relations among macronodes are expressed using rhetorical relations (Mann and Thompson 1987). A rhetorical link specifies the coherence relation existing between two portions of discourse, for example, whether text span B describes the event that caused state A to occur (CAUSE rhetorical relation) or whether text span B provides background information useful for interpreting the assertion in text span A (BACKGROUND relation). Rhetorical relations can also be used to represent coherence among content in different modalities (André and Rist 1993).

The *MACRONODE formalism* has been designed for building a new kind of flexible hypermedia system. Data can be used both to reason on the deep semantic structure of the ongoing presentation and to adjust its surface form (for example, its linguistic realization).

The *MACRONODESERVER* is a rule-based system able to select the most appropriate sequence of macronodes for a given communicative goal. It decides what information to include in the presentation as well as the most suitable discourse structure on the basis of a model of the user and knowledge about the interaction. Rules check the rhetorical links between macronodes (see a sample in the figure) and guarantee the global coherence of a presentation (for more information, see Not and Zancanaro [1999]).

As an example, let us compare the two sample texts here that were generated by the HIPS (hyperinteraction within the physical space) system.

The two texts provide a description of the same painting for two different visitors. Example 1 is a description built using an *elaboration strategy:* There are a lot of details, and the description is quite long. Example 2 is a description built using a *comparison strategy:* The text is rather short, and the description is mainly given by comparison with another fresco. One can note the slightly different beginnings of the two presentations ("This is the great fresco…" and "In front of you, you can admire the great fresco …") depending on two different realizations of the same macronode.

Example 1: Elaboration-Based Description of La Maestà

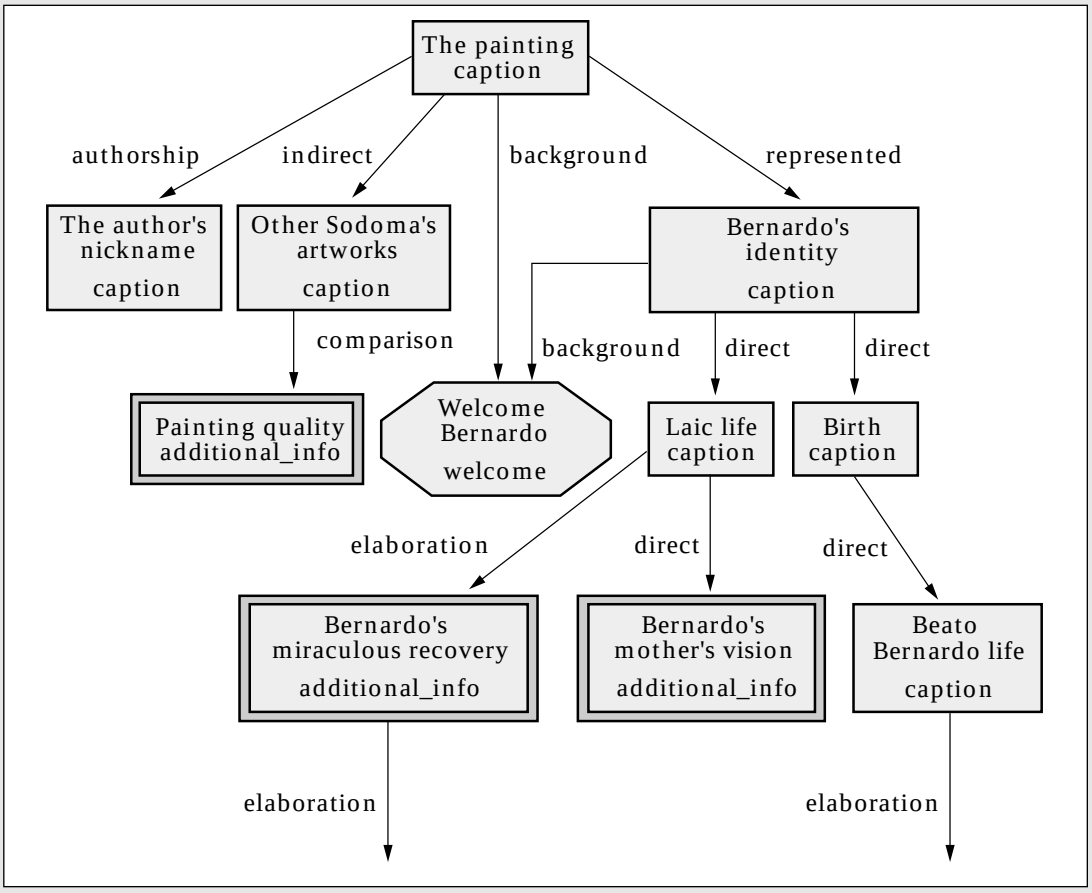
This is the great fresco La Maestà, depicted by Simone Martini in 1315. La Maestà was the first decoration piece of Palazzo Pubblico, therefore it acquired a particular value for the Sienese population through the centuries. It's not surprising that the very first decoration of Palazzo Pubblico (the site of political power) was a religious artwork. Only four years earlier, in fact, the 'Fabbrica del Duomo', the other civic power of Siena, influenced by the bishop, commissioned the famous 'Maestà' to Duccio di Boninsegna. The traditional spirit of competition between the two great 'factories' of the city demanded an adequate reply

Example 2: Comparison-Based Description of La Maestà

In front of you, you can admire the great fresco La Maestà, depicted by

concept yet, but we have worked "in vitro" on its development. We have considered the multimodal interface as a place where user and system actions occur that might be considered both as domain actions and communicative (linguistic and nonlinguistic) actions. In particular, when the system holds the initiative, it performs some domain actions, some communicative actions, or actions of both kind. At the same time, the interface is a sensorial organ, the collection of media through which the message is realized, and the (virtual) place where domain actions are actually performed. In the old teletype approach, described earlier, this ambiguity was not present.

The intentional structure of discourse has been modeled in Lochbaum (1998). Her proposal emphasizes the collaborative aspect of communication by means of a peculiar kind of plan called *SharedPlans.* The theory of SharedPlans (Grosz et al. 1999; Grosz and Kraus 1996) is based on the notion of plans as complex mental attitudes in which emphasis is put on the difference between the plans that an agent knows (that is, recipes for actions) and the plans that an agent adopts (that is, a structured collection of beliefs and intentions). The SharedPlan theory is intended to model interaction as a joint activity in which the participants try to build a plan together (in the second way of the previous distinction): The plan is shared in the sense that participants have a compatible set of beliefs and intentions. In this framework, communication is seen as the way



A Sample Portion of a Macronode Network.

Simone Martini in 1315. The fresco is located in the main part of the hall, the central point that gave the orientation of the Sala del Mappamondo. By contrast the Guidoriccio fresco, on the opposite side of the room, was a sort of great 'poster', glorifying the power of the Siena Republic. It was a sort of historical documentation more than an artwork to be judged for its artistic value.

in which agents agree on the various stages of the plan construction.

The difficulties in applying the SharedPlan theory to multimodal interaction (see also Rich and Sidner [1998]) arise from the double nature of the interface: Some actions (especially the linguistic ones) are intended to augment the current SharedPlan, but others are primarily intended to execute the related recipe; however, at the same time, if these actions take place on the interface, they also contribute to the augmentation of the plan. For example, if an agent is committed to do an action, it must perform the action and then inform the other agent of its execution. However, if the effects of the action are apparent on the interface, neither the explicit commitment nor the informing are actually necessary.

Any intelligent multimedia system requires a component that exploits the context to make presentation decisions (media selection, coordination, allocation, and so on) or interpret multichannel input (Maybury and Wahlster 1998). In particular, given information that needs to be displayed to the user, a multimedia coordinator automatically builds a coherent and coordinated presentation using a combination of available media (see, for example, Wahlster et al. [1992]).

Following Arens, Hovy, and Vosser (1993), any complex multimedia coordinator needs to be built around a collection of models: a model of virtual devices; a model of the characteristics of information to be displayed; a model

of the discourse and the communicative context; and a model of the interaction participants' beliefs, goals, attitudes, capabilities, and interests. Input and output processes interact with the dialogue manager that maintains the discourse structure and ensures a coherent interaction between the participants.

An important point is whether action execution is observable (and, in principle, interpretable as desired) by the other agent on the interface. Whether the action execution is observable depends on the ability of the multimedia coordinator to plan a meaningful presentation with the available media. The multimedia coordinator is instructed by the dialogue manager about the communicative intentions and returns the planned presentation to the dialogue manager. The dialogue manager, in turn, evaluates the expected effects on the other agent and whether the case asks for further planning. For example, if the presentation is not perspicuous enough, the dialogue manager might decide to plan a further communicative action (for example, an inform action).

We have has proposed a specific augmentation and execution process for SharedPlans that can accommodate this view (Zancanaro, Stock, and Strapparava 1997a). Two basic elements needed to find their place: (1) a local coherence technique that could be combined with the higher-level coherence of the SharedPlan approach that views communication as a collaborative activity and (2) multimedia coordination.

In explorative information access, it is more difficult for the system to recognize the user's intentions as far as real-world actions are concerned. The attentional aspect is more relevant, yet the intentional aspect can fruitfully be inserted. General strategies of exploration can be conceived, even if not every action on the part of the user can be interpreted at the planning level. Besides, some interaction fragments certainly can just be modeled as task oriented. A flexible combination of a more local-type representation and a collaboration-based one can be appropriate.

Conclusions

Language processing has a large practical potential if inserted in a multimodal conception of the interface. There are different dimensions for the concept of multimodality. One refers to the perception of different coordinated media used in delivering a message, another one to the combination of various attitudes in relation to communication and information access (for example, goal oriented and exploration oriented).

In this article, I took a practical perspective, and I referred to some implemented prototypes, mostly conceived for cultural tourism, a sector that I believe has a large potential. We have started with a system, developed some years ago, in which interaction was based on the seamless combination of navigation and dialogue. We have begun to take into account the physical space, with the goal of producing a personal, mobile device for person-oriented guided visits in a physical museum or a town. Toward the end, I also discussed introducing a deeper level of modeling multimodal dialogues based on collaboration.

As a final note about the application domain I discussed throughout this article, intelligent interfaces and language technology can help realize individual-oriented cultural tourism, yielding a more active role for the tourist. Other AI techniques look promising, for example, (1) case-based reasoning (Leake 1996), for adapting solutions that have proved successful for a similar tourist combined with dynamic information presentations, or (2) planning and temporal and spatial reasoning combined with advice about where to go during a visit, given a set of constraints (weather, crowd or traffic, terrain, interests, physical conditions, available time, and so on) or mechanisms for organizing the visit as an interactive game.

Increasingly, tourists will enjoy visits because they are culturally fulfilling experiences, conducive to learning and leading to further interests. Personal intelligent interfaces will be important elements for making the experience possible for all individuals.

Acknowledgments

I want to acknowledge the contribution of all the people at IRST who have worked on the ALFRESCO project—the list would be long—and on the HYPERAUDIO and HIPS projects (mainly Elena Not, Daniela Petrelli, Carlo Strapparava, and Massimo Zancanaro) with whom the ideas presented here were developed. Zancanaro and Strapparava were also essential for developing the collaboration-based multimodal dialogue work. I would also like to thank IRST's partners in HIPS, especially the colleagues from Siena University and Edinburgh University.

Notes

1. This article is an extension of an invited talk presented at the Sixteenth International Joint Conference on Artificial Intelligence.

2. The HIPS consortium includes University of Sienna (Italy, coordinating partner), CB&J (France), GMD (Germany), IRST (Italy), SIETTE-Alcatel (Italy), SINTEF (Norway), University of Dublin, and University of Edinburgh.

References

André, E., and Rist, T. 1993. The Design of Illustrated Documents as a Planning Task. In *Intelligent Multimedia Interfaces*, ed. M. Maybury, 94–116. Menlo Park, Calif.: AAAI Press.

Arens, Y.; Hovy, E.; and Vosser, M. 1993. On the Knowledge Underlying Multimedia Presentations. In *Intelligent Multimedia Interfaces*, ed. M. Maybury, 280–306. Menlo Park, Calif.: AAAI Press.

Benelli, G.; Bianchi, A.; Marti, P.; Not, E.; and Sennati, D. 1999. HIPS: Hyper-Interaction within the Physical Space. In Proceedings of IEEE Multimedia System '99, International Conference on Multimedia Computing and Systems, 1075–1078. Washington, D.C.: IEEE Computer Society.

Burgard, W.; Cremers, A. B.; Tox, D.; Hähnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; and Thrun, S. 1998. The Interactive Museum Tour-Guide Robot. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, 11–18. Menlo Park, Calif.: American Association for Artificial Intelligence.

Interface for Tourism. In *Information and Communication Technologies in Tourism 1999. Proceedings of ENTER'99*, eds. D. Buhalis and W. Schertler, 181–200. Vienna: Springer Verlag.

Grosz, B., and Kraus, S. 1996. Collaborative Plans for Complex Group Action. *Artificial Intelligence* 86(2): 269–357.

Grosz, B.; Hunsberger, L.; and Kraus, S. 1999. Planning and Acting Together. *AI Magazine* 20(4): 23–34.

Höllerer, T.; Feiner, S.; and Pavlik, J. 1999. Situated Documentaries: Embedding Multimedia Presentations in the Real World. Paper presented at the Third International Symposium on Wearable Computers (ISWC'99), 18–19 October, San Francisco, California.

Leake, D., ed. 1996. *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, Calif.: AAAI Press.

Lochbaum, K. 1998. A Collaborative Planning Model of Intentional Structure. *Computational Linguistics* 24(4): 525–572.

Mann, W. C., and Thompson, S. 1987. Rhetorical Structure Theory: A Theory of Text Organization. In *The Structure of Discourse*, ed. L. Polanyi. Greenwich, Conn.: Ablex.

Marti, P.; Rizzo, A.; Petroni, L.; Tozzi, G.; and Diligenti, M. 1999. Adapting the Museum: A Nonintrusive User Modeling Approach. Paper presented at the Seventh International Conference on User Modeling (UM99), 20–24 June, Banff, Canada.

Maybury, M. T., ed. 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park, Calif.: AAAI Press.

Maybury, M. T., ed. 1993. *Intelligent Multimedia Interfaces*. Menlo Park, Calif.: AAAI Press.

Maybury, M. T., and Wahlster, W., eds. 1998. *Readings in Intelligent User Interfaces*. San Francisco, Calif.: Morgan Kaufmann.

Mellish, C.; Oberlander, J.; O'Donnell, M.; and Knott, A. 1997. Exploring a Gallery with Intelligent Labels. Paper presented at the Fourth International Conference on Hypermedia and Interactivity in Museums (ICHIM97), 1–5 September, Paris, France.

Milosavljevic, M.; Tulloch, A.; and Dale, R. 1996. Text Generation in a Dynamic Hypertext Environment. Paper presented at the Nineteenth Australasian Computer Science Conference, 31 January–2 February, Melbourne, Australia.

Not, E., and Zancanaro, M. 1999. Reusing Information Repositories for Flexibly Generating Adaptive Presentations. In Proceedings of the IEEE International Conference on Information, Intelligence, and Systems, 566–569. Washington D.C.: IEEE Computer Society.

Not, E.; Petrelli, D.; Sarini, M.; Stock, O.; Strapparava, C.; and Zancanaro, M. 1998. Hypernavigation in the Physical Space: Adapting Presentations to the User and to the Situational Context. *The New Review of Hypermedia and Multimedia, Volume 4*, 33–46. London: Taylor Graham.

Rich, C., and Sidner, C. 1988. COLLAGEN: A Collaboration Manager for Software Interface Agents. *User Modeling and User-Adapted Interaction* 8(3–4): 315–350.

Stock, O. 1995. A Third Modality of Natural Language? In *Artificial Intelligence Review, Volume 2–3*, 123–146. Dordrecht, The Netherlands: Kluwer Academic.

Stock, O. 1991. Natural Language and Exploration of an Information Space: The ALFRESCO Interactive System. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, 372–378. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Stock, O.; Strapparava, C.; and Zancanaro, M. 1997. Explorations in an Environment for Natural Language Multimodal Information Access. In *Intelligent Multimedia Information Retrieval*, 381–388. Menlo Park, Calif.: AAAI Press.

Wahlster, W.; Andrè, E.; Bandyopadyay, S.; Graf, W.; and Rist, T. 1992. WIP: The Coordinated Generation of Multimodal Presentations from a Common Representation. In *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, eds. A. Ortony, J. Slack, and O. Stock, 121–144. New York: Springer Verlag.

Waterworth, J. H., and Chignell, M. H. 1991. A Model for Information Exploration. *Hypermedia* 3(1): 35–58.

Zancanaro, M.; Stock, O.; and Strapparava, C. 1997a. A Discussion on Augmenting and Executing SHAREDPLANS for Multimodal Communication. Paper presented at the American Association for Artificial Intelligence Fall Symposium on Communicative Action in Humans and Machines, 7–9 November, Cambridge, Massachusetts.

Zancanaro, M.; Stock, O.; and Strapparava, C. 1997b. Multimodal Interaction for Information Access: Exploiting Cohesion. *Computational Intelligence* 13(4); 439–464.

Zancanaro, M.; Stock, O.; and Strapparava, C. 1993. Dialogue Cohesion Sharing and Adjusting in a Multimodal Interactive Environment. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1230–1236. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Oliviero Stock is currently the director of the Institute for Scientific and Technological Research (ITC-IRST) in Trento, Italy. He joined IRST in 1988 to establish the natural language processing unit and has been responsible for the Cognitive Technologies and Communication Division since 1993. Until 1987, he worked at the Italian National Council for Research in Rome. His research activity has always been in AI, natural language processing, intelligent interfaces, and cognitive sciences and technologies. He is the author of more than 100 published papers and an editor of six books. He has been chairman of the European Coordinating Committee for AI (ECCAI) (1992–1994), president of the Italian Association for AI (AI*IA) (1992–1993 and 1994–1995), and president of the Association for Computational Linguistics (1996). He is an ECCAI fellow; has been on the program committee of some 50 international conferences and an invited speaker at about 30, including ECAI-92 and IJCAI-99; and has been on the editorial board of 10 scientific journals, including the *Journal of Artificial Intelligence* and *Computational Linguistics*, and associate editor of *Applied Artificial Intelligence*. His e-mail address is stock@itc.it.

Natural Language Processing and Knowledge Representation

Language for Knowledge and Knowledge for Language

Edited by Lucja M. Iwańska and Stuart C. Shapiro

Natural language refers to human language—complex, irregular, diverse, with all its philosophical problems of meaning and context. Setting a new direction in AI research, this book explores the development of knowledge representation and reasoning systems that simulation the role of natural language in human information and knowledge processing. As this book shows, the computational nature of representation and inference in natural language makes it the ideal level for all tasks in an intelligent computer system. The essays in this interdisciplinary book cover a range of implementations and designs, from formal computational models to large-scale natural language processing systems.

464 pp., illus., index ISBN 0-262-59021-2

PUBLISHED BY THE AAAI PRESS • COPUBLISHED BY THE MIT PRESS

Five Cambridge Center, Cambridge, Massachusetts 02142 USA
http://mitpress.edu/ • 617-625-8569 • 800-356-0343