Was the Title of This Talk Generated Automatically? Prospects on Intelligent Interfaces and Language

Oliviero Stock
ITC-IRST
38050 Povo, Trento, Italy
stock@irst.itc.it

Abstract

We are beginning to make use of technology that intervenes in the contents of the communication. Language processing has indeed a large practical potential if we take into account multiple modalities of communication. Multimodality refers to the perception of different co-ordinated media used in delivering a message but also to the combination of various attitudes in relation to communication and information access (e.g. goal-oriented and exploration-oriented). In the paper reference is made to some prototypes developed at IRST, conceived for cultural tourism. In a recent one the specificity is the combination of two forms of navigation taking place at the same time - one in information space, the other in the physical space. Some challenges for the future are discussed toward the end.

1. Introduction

Let me take it from far away. Language is the extraordinary means provided by the human mind for communicating with other humans (and for structuring thoughts). Spoken language has been for a long time 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. At the beginning it was pictorial, after a few centuries cuneiform coding was proposed. Only much later, in the 13th century B.C. according to some archaeologists, was an alphabet first introduced (in Ugarit, in what now is Syria). It consisted of about 30 cuneiform signs and was, so to say, quickly recognised as a breakthrough and was adopted by several peoples.

Meanwhile means for producing various instances of a written "document" were invented soon after the first appearance of written language. They consisted of cylinders with engravings (that could leave clay tablets impressed with text). It took more than 4000 years to get to Gutenberg and his flexible printing system, based on the alphabet.

It took 500 years more to get to the computer and its possibilities. Shortly before that, some other means for long distance communication (e.g. the telephone or the telex) also appeared, and in being adopted have produced some sort of slight variation of the basic two modalities (the spoken and the written ones).

With the computer the flexibility in dealing with the form of written language (editing) and accessing language (retrieving) is extremely emphasised. But in the scientific area of natural language processing, the goal has been 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 in making it possible to interact with a computer, in order to get some desired information. Though results have been fairly significant, the impact has been so far scarce, if we consider the way in which society could change if computers could really take the burden of understanding human language and make information available to all people. It took quite a lot for the scientific community to understand that communicating with a computer through natural language may mean something different from the two basic language modalities we are used to [Maybury and Wahlster, 1998]. The so-called teletype approach has persisted for some time before we began to understand the fact that a larger bandwidth of communication can be established between human and computer. For instance language can be integrated with images dynamically; the screen itself is not only an output medium but it 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].

But the point is not only in the interface. Often the

user does not know what information is available to her, or she may not have a clear idea of what she is searching for. The need arises for systems that integrate a mediated information access paradigm and a navigational paradigm, where the user may use different modalities to explore the material. We believe that exploration of an information space will become more and more a typical interactive attitude on the part of users, in a world inhabited by a multitude of available multimedia information [Maybury, 1997]. All this 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 of having a model of the user, including her interests, idiosyncrasies and the dynamic aspects inferred during the interaction. This is instrumental for making sense of partial or not detailed requests (or other acts) by the user, and for determining the system's actions.

A desirable feature is creating the appropriate presentations of information: 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 favoured.

If information is to be presented in a flexible way, it is essential that an automatic processor does the job – in the case of text presentation, a natural language generation system. The latter is a computational tool that automatically "builds" a text (a sequence of sentences) starting from abstract (non-linguistic) specifications. Given the internal representation of the knowledge sources, the system decides what is the relevant information to be communicated, it organises 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 organise media allocation and media co-ordination.

But our goal will be only partially attained if we do not begin to touch on the most difficult challenge of all: the challenge of keeping attention high, of building a seductive interface, of having the user be surprised and attracted by the creative attitude of the (artificial) companion.... At the end we are very rapidly getting to the point where whatever message, whatever agent, whatever interface will have to compete over the most precious and scarce resources human beings have: time, attention and emotional involvement.

I will discuss in turn some of these themes, making reference to work we have developed at IRST.

2. Multimodality and exploration of information

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

Fig. 1 A representation of ambiguity in multimodal communication (The Calling of St. Matthew, by Caravaggio, Rome, 1600)

Multimodality can also refer to the combination of various attitudes in relation to communication and information access (e.g. goal-oriented and exploration-oriented), each one having in isolation its specific characteristics.

Another view is concerned with the relation to human capabilities. Multimodality may amplify our capabilities. In order to do so it must be cognitively compatible, but altogether it may not reflect our "natural" communication.

Yet another aspect is related to the specific role that language plays in the multimodal context [Stock, 1995]. It is a different role than when communication is based on a single modality, and so its operational characteristics are different; its modelling requires specific components.

Integration of NLP 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: structural responsibility and target orientation. Structural responsibility involves the issue of which agent (i.e. the user or the system) is responsible for carrying out search and for 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 her actions nor by the configuration of the system. In reality there is a continuum of user behaviours varying between querying and browsing so that 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 then had to include a communicative action co-ordinator, responsible for proper media usage (and so, for example, it can take into account the deictic context at any time of the interaction) and for suggesting to the user a shift along the structural responsibility dimension.

Let me now summarise a system called AlFresco. For this system and for some other experiences I report here you will see a common application theme: cultural heritage and tourism. No wonder: Italy is considered to have half of the world's cultural tourism resources. Besides, the field can provide a wonderful opportunity for introducing technology that can help shifting from a mass-oriented attitude to an individual-oriented attitude – exactly what we aim at. Cultural tourism can become an experience where the individual is the active subject of the exploration, one who develops a personal taste and interest.

AlFresco [Stock, 1991] is an interactive, natural-language centred system for a user interested in Fourteenth Century Italian frescoes. It has the aim of providing information, and also of promoting other masterpieces that may 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 by manipulating various entities, images, and text itself. The system builds a simple model of the user as the dialogue proceeds and uses it for output decisions, while 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 may cause zooming into details or changing the focus of attention onto other frescoes.

We have proposed a level of multimodal acts representation [Stock *et al*, 1997], roughly corresponding to what, for strictly linguistic dialogues, is the illocutionary level (Fig. 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.

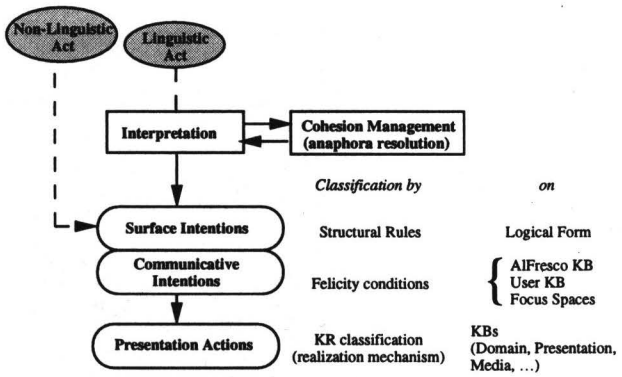

Fig. 2 AlFresco dialogue management

The dialogue cohesion management [Zancanaro *et al*, 1993, 1997a] provides also a graphical feedback of the dialogue cohesion status to the user. This visual representation: (a) reassures the user at a glance on the system's interpretation (as such it takes the place of a paraphraser), and (b) allows co-operative recovery from

discourse misconceptions by means of a series of "intuitive actions" when this interpretation is not the one the user meant.

In general a tighter integration of different modes of exploration, language-oriented and navigational, has been accomplished. I think this is a very fruitful concept that 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 understand better this kind of amplification of human communicative capabilities.

3. Bringing physical space into the picture

One further element for advancing in the direction of personalisation 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 in moving in a physical space and acquiring information about the objects shown (and of course in 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 centre of the physical-virtual space exploration and 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,* to appear]. 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, a European project of the Esprit I^3 program[1].

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 [e.g. ILEX, Mellish *et al* 1997] or dynamic encyclopaedias [e.g. PEBA-II, Milosavljevic *et al*, 1996]. Moving in a physical museum has been the goal of the RHINO project, where a robot accompanies the visitor [Burgard *et al*, 1998].

Content adaptation in a physical environment poses some problems that are related to the fact that the visitor is experiencing a "real" situation: the cognitive problems that may arise when a person is moving in a virtual information space, are different when the user is seeking concrete objects, moving in a real environment that provides stimuli, attention grasping and feedback. Information is presented in different situational contexts, determined mainly by: (i) user position and movements; (ii) the structure of the surrounding physical space (e.g., whether objects are close or not); (iii) whether other people are examining the same item or not; (iv) whether the user came alone or not.

HyperAudio (and HIPS) integrates the individual, dynamic modelling of the user with a general model of the environment, of the user's movements and of the discourse history to best tailor information presentations. Different 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 rethorically 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 via a 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 (Fig. 3).

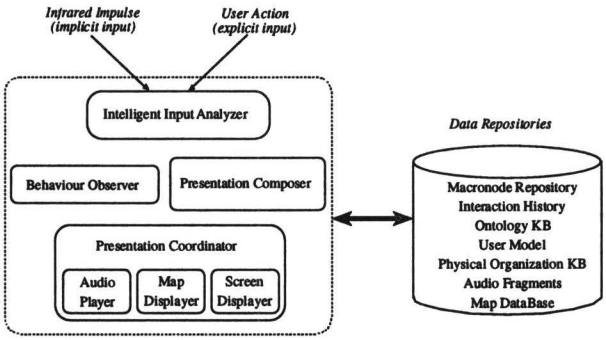

Fig. 3 Architecture

The user model is accessed to exploit: a) the user's interests (which are inferred from her behaviour), to include in the new message information that can stimulate the hearer's attention, possibly proposing information about other objects/sites strictly related to what the user is seeing, to increase curiosity and desire to explore; b) the user's background knowledge, in order to relate the new information presented to what she already knows (therefore reinforcing learning) and to

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

decide whether additional clarification or exemplification of new concepts is required to help her understand. As the interaction proceeds, the system refines its assumptions about the user's interests and knowledge by observing the user's behaviour and by keeping track of the information she has been exposed to. 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 realise part of the content as clickable links on the screen to avoid overwhelming the visitor with information.

The language style adopted for each single user is selected according to the user type.

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

The system includes also a graphical interface that helps to orientate the visitor and is useful for complementing linguistic instructions. Besides, "clickable" elements in the oral presentations appear on the screen.

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 analysed by the Intelligent Input Analyser that decides on the most suitable type of processing required (e.g., plan a new presentation, stop the current presentation, plan a navigation support message, etc.). From all input the Behaviour 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. 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 (e.g., introductory label, 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 realise the content of the macronode and thus encodes its surface linguistic forms, while the relevant concept, the rhetorical links and the message type encode the deep structure of the description.

Many of the issues presented for the museum setting apply to any physical hypernavigation setting in which individual, dynamic guides would be appropriate: for example historical 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 GPS as the localisation system) and suggest more ambitious scenarios (for example: new functions to support groups of visitors, access to on-line services such as meteo forecasts). All that is subject to investigation inside the European I^3 Project HIPS. 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: there is a basic path where the visitor, through presentations and suggestions is attracted, whatever deviations she performs; distances from a position to the next position tend to be minimised. Another one brings into the picture the dimension of play: for instance a dynamic treasure hunt, where typically physical distances tend to be maximised.

Yet another innovative feature is the introduction of collective memories. The visit trace is kept. The data can be used by the visitor: when she is back home she will be able to go deeper into the domain she has been exposed to, with a system (like AlFresco) that knows about her visit and will support her in her successive exploration.

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

3.1. What about speech?

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

Coming to spoken input, there has been a lot of progress in spontaneous speech recognition, albeit in highly constrained dialogic settings. For instance within the C-STAR II consortium IRST and its partners have built a prototype aimed at making possible that two persons physically remote and each speaking her own language [Cettolo *et al*, 1999], 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, hesitations and so on. Recently we have begun to work on spoken dialogue within a highly sophisticated, immersive, graphical system about cultural heritage.

4. A challenge - Collaboration

What 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 concept yet, but we have worked "in vitro" on this. We have considered the multimodal interface as a place where actions occur that may be considered both as domain actions and communicative (linguistic and non-linguistic) actions. This is true for user and system actions: when the system holds the initiative it performs some domain actions, some communicative actions or actions of both kinds. The interface is at the same time a sensorial organ, the collection of media through which the message is realised, and the (virtual) place where domain actions are actually performed. In the old "teletype approach" this ambiguity was not present.

The intentional structure of discourse has been modelled in [Lochbaum, 1999]. Her proposal emphasises the collaborative aspect of communication, by means of a peculiar kind of plans called SharedPlans. The theory of SharedPlans [Grosz and Kraus, 1996] is based on the notion of "plans as complex mental attitudes", and is intended to model interaction as a joint activity in which the participants try to build a plan together: 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 in which agents agree on the various stages of the plan construction.

The difficulties in applying the SharedPlan theory to multimodal interaction arise from the double nature of the interface: some actions (especially the linguistic ones) are intended to augment the current SharedPlan while others are primarily intended to execute the related recipe, but at the same time, if these actions take place on the interface, in some way they contribute to the augmentation of the plan too. For example, if an agent is committed to do an action it must perform it and then inform the other agent of its execution: but 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, co-ordination, allocation, etc.) or to interpret multi-channel input [Maybury and Wahlster, 1998]. In particular, given information that needs to be displayed to the user, a multimedia co-ordinator automatically builds a coherent and co-ordinated presentation using a combination of available media. [see for intance Wahlster *et al*, 1992].

Following [Arens, 1993] any complex multimedia co-ordinator 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, 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. This depends on the ability of the multimedia co-ordinator to plan a meaningful presentation with the available media. The multimedia co-ordinator is instructed by the dialogue manager as to 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 instance in case the presentation is not perspicuous enough, it may decide to plan a further communicative action (for example an inform action).

We have proposed a specific augmentation and execution process for SharedPlans that can accommodate our view [Zancanaro *et al*, 1997b]. Two basic elements needed to find their place: a) 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 b) multimedia co-ordination.

In explorative information access it is more difficult for the system to recognise the user's intentions, as far as real world actions are concerned. The attentional aspect is more relevant; yet the intentional aspect can be fruitfully inserted as well. 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 modelled as task-oriented. A flexible combination of a more "localist" representation and a collaboration-based one can be appropriate.

5. Computational humour - Not so crazy

Interfaces must be seductive: they must attract and satisfy the user. This is particularly true of situations where the goal is not so much work productivity, but a reality that is a mix of entertainment, information and education: so called *edutainment* constitutes an increasing large portion of what the computer can offer to our life. We know a major challenge for our society will be to find ways to improve education, not only institutionally, but at all level of activity. And we know that learning can go well together with active entertainment.

The human mind likes communication and emotions. Personally I consider humour an essential part of communication. The relaxation of inner censorship and the release of energy that derives from it produce an intense pleasure that tends to repeat itself in a favourable situation for giving raise to this phenomenon. Interactive and individual-oriented computational humour [Stock, 1996], beyond entertainment, *per se* important, will help all kinds of concept promotion and, in general, of learning. I am sure, among application areas, it will constitute the key to a number of children's activity and games: think of the children's ability in finding ambiguities and absurd meanings and the match they could find in a computer; simple humour on the part of the system can be a great resource as it helps to memorise errors and corrections; it will help develop social behaviour, etc. Even in a domain of active exploration, humour can help keep attention high, promote items and memorise what is seen.

While humour without restrictions is certainly "AI-complete", there is work that shows the feasibility of introducing some elements in a system [see for instance Binsted 1996].

A restrictive view of interfaces, normal today, will eventually yield to a more advanced view, in which interaction with the user will be accomplished for instance through an assistant, a critic or through a group of characters. The work of Barbara Hayes-Roth [see for instance 1996] and others has shown some possible ways. Humour, I believe, has to play an essential role. The characteristics of the interface will determine its potential on education and on society in general and we should not miss this opportunity. It is worth remembering what Groucho Marx said of TV: "I find TV very educational. The minute somebody turns it on, I go to the library and read a good book."

Conclusions

Language processing has indeed 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 co-ordinated media used in delivering a message; another one to the combination of various attitudes in relation to communication and information access (e.g. goal-oriented and exploration-oriented).

In this presentation I have taken a practical perspective and I have 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, where interaction was based on the seamless combination of navigation and dialogue. We have then 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 have talked about bringing into the picture a deeper level of modelling multimodal dialogues, based on collaboration, and briefly discussed the relevance of seductive interfaces and humour.

What else do I want to say about prospects for realising more useful and intelligent interfaces?

- It is important to introduce elements of a cognitively compatible, qualitative physics in our systems, so that in our multimodal interaction the system can understand what the user's representations and reasoning about space and time are and how she naturally conveys meanings in her communication.
- Case-based reasoning techniques can be a very relevant resource for presentation systems, especially in the edutainement and cultural heritage domain. We must find appropriate ways of using this resource in the multimodal context.
- We need to make experiments, especially Wizard of Oz simulations and understand more about what is really cognitively appropriate. Most of the things we foresee do not exactly correspond to what we find in nature (face to face communication) or in written text. Our technology must not be intrusive, and on the other hand it may be different from other modalities we have known for a long time.
- Natural language generation has a large potential for improvement. One thing our community can yield is the next step after hypertext. A situation where text fragments are tailored and the reader/hearer may want to interact with the system so that the material is presented to her with a certain style, taking into account an increasingly large number of personal characteristics and behaviours. On our future radiosets, we shall have not only the loudness setting but also an interactive adaptation of the presentation to our context.

By the way, the answer to the question in the title is no. But the question may have a different answer when posed again in the not so far future.

Acknowledgments

I wish to acknowledge the contribution of all the people at IRST who have worked at the AlFresco, HyperAudio,

and HIPS projects, with whom the ideas presented here have been developed. In particular, Massimo Zancanaro and Carlo Strapparava, beside all this, have also been essential for developing the collaboration-based multimodal dialogue work.

References

[Arens *et al*, 1993] Y. Arens, E. Hovy and M. Vosser. On the Knowledge Underlying Multimedia Presentations. In M.T. Maybury (ed.) *Intelligent Multimodal Interfaces*. AAAI-Press/MIT Press, Menlo Park CA/Cambridge MA, 1993.

[Binsted, 1996] K. Binsted. Machine Humour: An Implemented Model of Puns. PhD Thesis. University of Edinburgh 1996.

[Burgard *et al*, 1998] W. Burgard et al. The Interactive Museum Tour-Guide Robot. *Proceedings of AAAI-98, Fifteenth National Conference on Artificial Intelligence*, Madison, 1998.

[Cettolo *et al*, 1999] M. Cettolo, A. Corazza, G. Lazzari, F. Pianesi, E. Pianta, L. M. Tovena. A Speech-to-Speech Translation-Based Interface for Tourism. In D. Buhalis, W. Schertler (eds.), *Information and Communication Technologies in Tourism 1999. Proceedings of ENTER'99*, Springer Verlag, Vienna, 1999.

[Grosz and Kraus, 1996]. B. Grosz and S. Kraus. Collaborative Plans for Complex Group Action. *Artificial Intelligence* , 86(2), 1996.

[Hayes-Roth, 1995] B. Hayes-Roth. Agents on Stage: Advancing the State of the Art of AI. *Proceedings of IJCAI-95, Fourteenth International Joint Conference on Artificial Intelligence.* Montreal, 1995.

[HIPS] HIPS Project, WWW home page: http://www.ing.unisi.it/lab_tel/hips/hips.html.

[Lochbaum, 1998] K. Lochbaum. A Collaborative Planning Model of Intentional Structure. *Computational Linguistics*, 24(4), 1998.

[Maybury, 1993] M.T. Maybury (ed.) *Intelligent Multimedia Interfaces*, AAAI Press, Menlo Park, Ca./MIT Press, Cambridge, Mass., 1993.

[Maybury, 1997] M.T. Maybury (ed.) *Intelligent Mutimodal Information Retrieval*, AAAI Press, Menlo Park, Ca./MIT Press, Cambridge, Mass., 1997.

[Maybury and Wahlster, 1998] M.T. Maybury and W. Wahlster (eds.) *Readings in Intelligent User Interfaces*, Morgan-Kaufmann Press, San Francisco, 1998.

[Mellish *et al*, 1997] C. Mellish, J. Oberlander, M. O'Donnell and A. Knott. Exploring a Gallery with Intelligent Labels. *Proceedings of the Fourth International Conference on Hypermedia and_Interactivity in Museums (ICHIM97)*, Paris, 1997.

[Milosavljevic *et al*, 1996] M. Milosavljevic, A. Tulloch and R. Dale. Text Generation in a Dynamic Hypertext Environment. *Proceedings of the Nineteenth Australasian Computer Science Conference*, Melbourne, 1996.

[Not *et al*, to appear] E. Not, D. Petrelli, M. Sarini, O. Stock, C. Strapparava, M. Zancanaro. Hypernavigation in the Physical Space: Adapting Presentations to the User and to the Situational Context. To appear in *The New Review of Hypermedia and Multimudia.*

[Stock, 1991] O. Stock. Natural Language and Exploration of an Information Space: the AlFresco Interactive System. *Proceedings of IJCAI-91, the Twelfth International Joint Conference on Artificial Intelligence*, Sydney, 1991. Also in M.T. Maybury and W. Wahlster (eds.) *Readings in Intelligent User Interfaces.*, Morgan-Kaufmann Press, San Francisco, 1998.

[Stock, 1995] O. Stock. A Third Modality of Natural Language? *Artificial Intelligence Review*, 9, Kluwer Academic Publishers, Dordrecht, 1995.

[Stock, 1996] O. Stock. Password Swordfish: Verbal Humour in the Interface. *Proceedings of IWCH, International Workshop on Computational Humour*, Enschede, 1996.

[Stock *et al*, 1997] O. Stock, C. Strapparava and M. Zancanaro. Explorations in an Environment for Natural Language Multimodal Information Access. In M. Maybury (ed.) *Intelligent Mutimodal Information Retrieval.* AAAI Press, Menlo Park, Ca./MIT Press, Cambridge, Mass., 1997.

[Wahlster *et al*, 1992] W. Wahlster, E. Andrè, S. Bandyopadyay, W. Graf and T. Rist.. Wip: The Coordinated Generation of Multimodal Presentations from a Common Representation. In A. Ortony, J. Slack, and O.Stock (eds.), *Communication from Artificial Intelligence Perspective: Theoretical and Applied Issues*, Springer Verlag, 1992.

[Waterworth and Chignell, 1991] J.H. Waterworth and M.H. Chignell. A Model for Information Exploration, *Hypermedia*, 3, 1991.

[Zancanaro *et al*, 1993] M. Zancanaro, O. Stock and C. Strapparava. Dialogue Cohesion Sharing and Adjusting in a Multimodal Interactive Environment. *Proceedings of IJCAI-93, Thirteenth International Joint Conference on Artificial Intelligence*, Chambery, 1993.

[Zancanaro *et al*, 1997a] M. Zancanaro, O. Stock and C. Strapparava. Multimodal Interaction for Information Access: Exploiting Cohesion. *Computational Intelligence*, 13(4), 1997.

[Zancanaro *et al*, 1997b] M. Zancanaro, O. Stock and C. Strapparava. A Discussion on Augmenting and Executing Sharedplans for Multimodal Communication. *Proceedings of the American Association for Artificial Intelligence Fall Symposium on Communicative Action in Humans and Machines*, Boston, 1997