Pragmatics and Human Factors for Intelligent Multimedia Presentation: A Synthesis and a Set of Principles

Frédéric Landragin

Pragmatics and Human Factors for Intelligent Multimedia Presentation: A Synthesis and a Set of Principles

– DRAFT VERSION –

Frédéric Landragin

CNRS, LaTTICe Laboratory (UMR 8094)
Montrouge and Paris, France.
Email: frederic.landragin@linguist.jussieu.fr.

Abstract

Intelligent multimedia presentation systems (IMMPS) have to take into account pragmatics and human factors such as the specificities of human perception, attention, memory, conceptualization, and language. Using a conversational animated agent or not, some principles can be followed to increase the communicative abilities of interactive systems. In this paper we propose a set of such principles. We exploit our background in natural language processing and computational pragmatics to provide specifications for multimodal systems. The classifications and principles and architectural concerns we present are based on some experimental observations (that are not described here) and constitute a kind of white paper for future implementations.

1 INTRODUCTION

An intelligent multimedia presentation system (IMMPS, see Bernsen [4], Bordegoni et al. [5], Karagiannidis et al. [12] and others) has to translate display requests from a dialogue manager into output messages, and therefore has to take into account the particular characteristics of the information, the terminal, the physical environment, and the addressee (or user). When information has to be spread over several communication channels that correspond to different communicative modalities, we talk about multimodal fission. The term 'information' groups natural language and multimodal utterances from the user and the system as well as the associated application data.

Following this definition that characterizes our approach, we can distinguish two parts for the presentation process in a dialogue system. First, the dialogue manager takes the **decisions** on the following aspects ('WH- part' of the process):

- '**Who**' = to whom the information has to be presented,
- '**What**' = what is the information to present,
- '**Which**' = which part of the information has to be emphasized,
- '**Where**' = where can the information be displayed, i.e., on which devices,
- '**When**' = when and for how long must the information be presented.

Second, the IMMPS **realizes** these decisions ('HOW part' of the process) by: choosing the method to valorize the related piece of information (cf. 'which' in the previous list), choosing the modality or modalities and the device or devices to exploit (cf. 'where'), dividing the information to determine the related pieces of information for each modality (cf. 'where'), dividing the information to spread its presentation over time (cf. 'when'), and, possibly but not necessarily, managing a human-machine interface (HMI), for instance a graphical user interface (GUI), that is specific to the presentation, e.g., navigation buttons when information has to be split for several display steps.

With these simple items, we want to make precise the roles of intelligent multimedia presentation. Our proposal is not that different from existing ones like [19] or others, but it includes as many aspects as possible, in particular a clear separation between the dialogue concerns and the presentation concerns. These items are valid whatever the form of the IMMPS (avatar or not), whatever its communicative status, from a fully recognized interlocutor to a simple intermediary with the application. More precisely, the IMMPS can have the status of an interlocutor, i.e., can stand as a 'majordomo'. The user can interact with it, the details of the exchanged information having no interest for the application or for the dialogue manager. The advantage is that the user's actions that concern only the HMI or GUI are treated very quickly. To the contrary, the IMMPS can have no materiality for the user, who believes he/she is communicating directly with the application. The advantage here is the simplicity and transparency for the user.

After a section presenting some first principles for taking into account pragmatics and human factors in multimedia presentation, we will focus on the determination of all input parameters that a presentation system may take into account. These parameters are presented in a set of classifications. A general architecture illustrates the processes that exploit them. These processes are grouped into two main steps that are then described in details, with examples of rules and strategies for multimedia presentation.

2 FIRST PRINCIPLES

2.1 Nine principles for IMMPS

Our approach and preoccupations can also be summarized into a set of principles, which can be compared to the Grice's maxims dealing with more general conversational principles [10]. To us, designing more natural and adaptive IMMPS requires that the characteristics of the information (or message) in its context, in particular the

linguistic context or dialogue history, are taken into account in a better way. This first point leads us to propose four principles:

1. “More natural IMMPS with a better repartition of information over the communication channels”,
2. “More natural IMMPS with a natural rendering and valorization of the information on a communication channel”,
3. “More natural IMMPS with a better exploitation of the semantic content of the message”,
4. “More natural IMMPS by maintaining better cohesion and coherence with previous messages”.

Second, designing more natural IMMPS (more natural in the sense of more used-centric, with more naturalness and adaptative abilities) requires taking into account the characteristics of the terminal (presentation means), and the physical and situational environment (presentation conditions). Hence, we propose the two following principles:

5. “More natural IMMPS with a more refined exploitation of presentation means”,
6. “More natural IMMPS with a more refined exploitation of presentation conditions”.

Third, to provide more user-oriented IMMPS, i.e., presentation systems that are more sensitive to human abilities and behaviors, there is a particular need to take into account the addressee’s physical and cognitive abilities, as well as his role(s) in the application domain and preferences for information presentation. Three additional principles can then be expressed:

7. “More natural IMMPS with a better exploitation of the addressee’s expectations”,
8. “More natural IMMPS for a better perception of the message by the addressee”,
9. “More natural IMMPS for more relevant future reactions from the addressee”.

2.2 Multimedia information

Multimedia information can have many forms and contents. Some characteristics are essential when presenting. Among them, we can cite following Arens and Hovy [2] the urgency (‘urgent’ or ‘routine’, for instance), the transience (‘live’ or ‘dead’), the critical importance or criticality (‘nominal’, ‘critical’, ‘fatal’), the density or structure (‘continuous’, ‘discrete’), the coverage or number of simple items that are grouped into one complex structure (‘singular’, ‘low’, ‘high’, ‘total’), the volume (‘much’, ‘little’, ‘single’) and so on. Moreover, the application often consists of managing complex information such as cartography or video. A lot of work also deals with the best ways to

represent such information with a particular concern on adaptation to the context. Still, choosing the relevant characteristics given a particular application remains a complex problem. We will try to extract the characteristics that are essential to our proposal from all the ones mentioned.

Another aspect of research work dealing with multimedia information is the representation of such information for communicative systems. A lot of standards or standard proposals have been designed: EMMA from the W3C [7, 25], SMIL that focuses on the synchronization problems [3, 26], MPML [17] and others. Even if studying such initiatives can provide ideas on how to represent multimodal information, our approach is at too early a phase to exploit them. Semantics, pragmatics and user's abilities are not the main preoccupations of these initiatives, but they are ours.

2.3 Human factors for IMMPS

Whereas the term 'adaptability' is used for the adaptation of the interface by the user and is studied at design time, 'adaptivity' is used for the adaptation of the interface by the system itself, at run-time. Then, adaptivity groups all dynamic aspects of adaptation and is very close to our concerns about IMMPS and human factors. More precisely, following work such as [8], we can state that:

- adaptation to the **terminal** intervenes during the presentation because the presentation method depends on the terminal characteristics,
- adaptation to the **physical environment** intervenes during the presentation because criteria such as background noise level consist of parameters for IMMPS,
- adaptation to the **user's preferences** intervenes during the presentation (it is of course in the interest of IMMPS to follow the display preferences),
- adaptation to the **user's roles** (or user task) intervenes during the presentation because IMMPS can exploit its knowledge of the user's roles for emphasizing a piece of information,
- adaptation to the **user's access rights** (or user's prerogatives or profile) intervenes before the presentation because the dialogue manager has to filter the information that the user must not know.

Information presentation and information adaptation are thus two similar problems. The important point is how information is really adapted to the user's preferences and abilities. In particular, cognitive abilities such as attention, perception, immediate memory, mental representation, conceptualization, judgment, decision, and so on, can have an influence on how information could be best represented. A lot of work has been done on communicative agents and avatars, for instance the current research on emotion rendering, but there is a lack of implementation of theories dealing for instance with the Gestalt Theory [13] or salience, for instance. What we want to address here is the integration of such factors into IMMPS in order to have a certain control on the user's behavior. For that, we will follow work such as [23] and we will extend our approach for natural language and multimodality understanding.

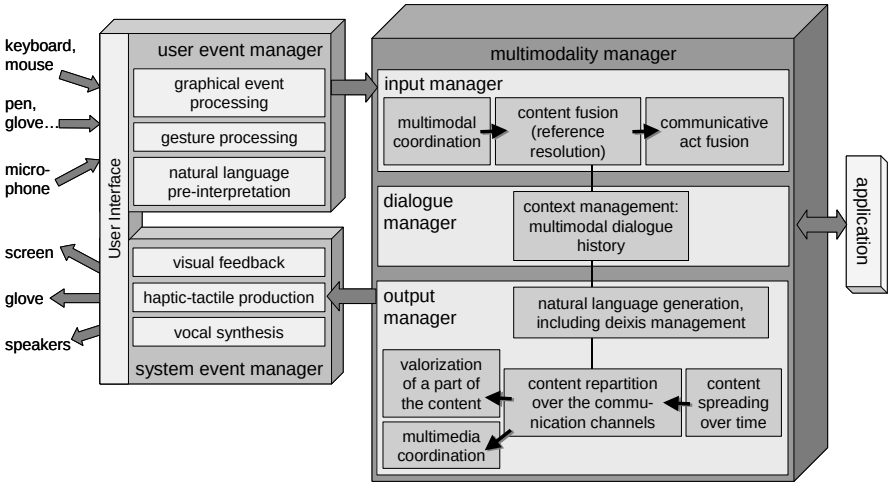

Figure 1: Architecture for input and output multimodality management

3 IMMPS AND HUMAN FACTORS

3.1 General presentation

Figure 1 presents the global architecture that underlies our approach. The core is the multimodality manager that treats input and output multimodality, and manages the multimodal context, i.e., the history of multimodal utterances and actions from the user and the system. The components of the output manager and the parameters they exploit are described in the following subsections.

3.2 Input parameters for IMMPS

Input parameters can be separated into three categories. The first relates to the information to be presented, and includes the structure and content of the information as well as the pragmatic force it is associated with. The second relates to the presentation means and groups the terminal and the physical environment. The third relates to the presentation addressee (the user) and groups the preferences, abilities, roles, and all human factors, with a sub-distinction between physiological factors, linguistic factors, and cognitive factors.

3.2.1 Information-related parameters

The criteria used for the repartition and valorization of the information and related to the information itself can be classified into three categories. The first deals with the message content and includes: (a) the level of criticality, (b) the level of urgency, very important because IMMPS must be able to stop any process when an urgent information has to be displayed, (c) the information complexity, i.e., some precisions about the data structuring and the size and numbers of items, (d) the information constitution,

i.e., some precisions about its density (discrete or continuous, list or table of items, timetable, etc.), (e) the information scope, for instance the fact that the information has two poles of interest, firstly the whole information and secondly a zoom in on one particular element, and (f) the presentation constraints that are inherent to the information: visual constraint for a cartography, no constraint for a linguistic message or for data that can either be displayed or verbalized. Note, the distinction between the level of criticality and the level of urgency is important because it has an influence on the IMMPS behavior. Critical information should be presented using a particular rendering to make it obvious to the user, whereas an urgency should stop all the current processes so that the user is face to face with it and only it. There is concerning the information scope an important aspect linked to the users' perceptive abilities. In the case of a huge table of numeric values as the information to present, two complementary strategies can be imagined. The aim can first be to present the information in its entirety, so that the user can apprehend its scope in one glance, even if no value can be read because of the very low font size that is required. A second aim can be to present the content of one particular cell to the user, and then to exploit a kind of magnifying glass. The method that groups both aims is sometimes called 'keeping the context'. With the information scope and the privileged aim as parameters, IMMPS should be able to choose between one strategy, the other, or both.

The second category of parameters groups pragmatic aspects with illocutionary and perlocutionary forces of the message: (a) the communicative act(s) that is/are determined by the dialogue manager, and (b) the expected reaction from the user: feedback or not, immediate action or not. Illocutionary and perlocutionary forces [21] will be described in detail in section 3.3.

The third category relates to the interaction history: (a) the history of the display actions, in order to allow the mention of a previously executed action, and (b) the stack of the displayed data, in order to allow the mention of previously displayed data.

3.2.2 Presentation means-related parameters

The characteristics of the terminal constitute a first set of parameters related to the presentation means: (a) terminal availability, (b) dimension constraints such as screen size, (c) constraints on the processing delays, and (d) constraints and preferences on output modalities.

A second set of parameters consists of the parameters that are related to the presentation environment. For these, we propose to exploit and adapt the information presentation to the three functions of gesture that were identified by Cadoz [6] for the gesture as an input in HMI: (a) **epistemic** constraints, that are linked to the 'learning from the environment' function, typically picking up and taking into account the ambient noise and the ambient luminosity, (b) **ergotic** constraints, that are linked to the 'transforming, changing the state of the environment' function, typically thresholds for ambient noise and luminosity, that must not be overstepped in order to not disturb the environment, and (c) **semiotic** constraints, that are related to the 'communicating meaningful information toward the environment' function, typically the quantity and quality of speech delivery, e.g., too loud or too fast considering the environment.

3.2.3 User-related parameters

Four categories can be distinguished here. First, the parameters that deal with the user's physical abilities: (a) constraints on the ways of working with communication channels, for instance due to a handicap, and (b) constraints and preferences on the exploitation levels of the communication channels, e.g., when the visual channel is already monopolized by another part of the ongoing task. Here, the auditory channel has a particular role, because it is sometimes the only possible modality to convey a message (the user may use his hands for the ongoing task and can only use speech to express something else). Second, the parameters related to the user's roles: (a) constraints on the access rights and bans that come from the 'user profile' (a user-related resource that is managed by the dialogue manager), and (b) constraints and preferences that come from the ongoing 'user task' (another resource managed by the system). Third, we can group all other individual preferences, particularly: (a) the preferences for linguistic terms and presentation metaphors, which were previously expressed by the user, and (b) the preferences on the dialogue management, which are detected and exploited dynamically by the dialogue manager, e.g.: the user prefers short answers to long ones; the user always prefers to conclude a sub-dialogue before going back to the main dialogue. In the last category we can group all other human factors that correspond to universal preferences, i.e., preferences that apply to everybody due to (a) human physiology, (b) linguistic abilities, and (c) cognitive abilities.

Physiologic preferences are first linked to the modality. Within the sound modality, two statements are of importance in particular for beep or horn messages (also called earcons). First, the stronger the sound is, the more powerful it is (but the more stressful it is). Second, high pitch is more strident than low pitch. Similar statements can be taken into account for visual modality. Following color theories [11], red is perceived much quicker than blue or yellow, and therefore is more often exploited for visual alerts. Blue can be perceived much easier in dark environments than in luminous environments. The center of the visual field that corresponds to the fovea is a privileged place. The notion of 'good form' from the Gestalt Theory [13], for instance a perfect and simple circle, is a privileged form. Moreover, salience and pregnance can be relevantly exploited whatever the modality. A salient element, i.e., an element that can be distinguished by singular properties (e.g., the only red element), is more easily perceived. A pregnant element, i.e., an element that has been the object of previous repetitions so that it impregnates the user's memory, is also more easily perceived. IMMPS should exploit such criteria to optimize its presentations considering the particularities of human perception.

Concerning linguistic particularities, simple statements can also be done at the different linguistic levels. At lexical and syntactic levels, IMMPS may keep the terms and syntactic constructions from the user, and may, in a general manner, use simple words and constructions. At semantic and pragmatic levels, the Grice's maxims [10] may be exploited when determining the message to generate. The risks of ambiguities could be minimized, for instance by avoiding anaphora when several potential antecedents are possible. With the same purpose, indirect and composite speech acts should be avoided. At a stylistic level, the informational or communicative structure [15] should be exploited in order to put one particular message element forward. This 'putting into

salience' or 'saliencing' process is done by choosing the relevant grammatical function, thematic role, theme, focus, etc. Coherence (generating a message with a logical link with previous ones) and cohesion (generating a message whose form is in direct continuity with the form of previous messages) should be exploited.

Concerning cognitive preferences, the particularities of lower cognitive processes (perception, attention, memory) and of upper cognitive processes (mental representation, judgment, decision) should be clarified and taken into account. Then, IMMPS should be aware of the size of short-term memory (from 5 to 7 independent items, see [16]), of selective and persistent attentions, etc. More precisely, a message can have the purpose of capturing selective attention (e.g., alerts) or to request an important amount of persistent attention for a thorough treatment (e.g., presentation of an important information). IMMPS must give no opportunity for selective attention to be diverted in various directions, and should provide time to the user's persistent attention. Moreover, each message leads to a representation process whose complexity depends on the complexity of the information in its canonical form. So IMMPS should stay inside reasonable limits. Some pieces of information require a judgment. So IMMPS should not multiply such pieces of information in the same presentation act. Because of their visual characteristics, some pieces of information have an influence on the actions that can be done on them. IMMPS should manage such affordances [9] in a relevant way. In a general manner, it can be very efficient to exploit all that has already worked well. For instance, if the system noticed that a particular visual form has a positive and efficient influence on the user, it may decide to use it again in similar situations.

3.2.4 Statement on inputs and outputs

The constraints and principles we have described can be summarized in the following process:

- From the applicative domain, the user task and user profile: (a) levels of criticality and urgency, (b) self-descriptive information (organized and quantified information), and (c) presentation constraints and preferences that are specific to the task or task type;
- Computed by the dialogue manager: (a) pragmatic forces and other labels on the message, for instance an emotion to render, (b) coherence and cohesion indications, (c) linguistic valorizations, and (d) constraints and preferences on linguistic terms and dialogue management;
- Determined by IMMPS on the basis of the constraints from the previous items: (a) information ordering (e.g., depending only on urgency levels), (b) method to dissociate an information into several presentation phases, (c) method to dissociate an information over the communication channels, (d) for each piece of information, level of valorization (e.g., depending only on criticality), (e) method to valorize a piece of information, and (f) method to exploit the preferences, in particular when they contradict each other.

3.3 Pragmatics for IMMPS

Following our approach, human-machine dialogue systems should be able to communicate with their users in a spontaneous and natural way, by exploiting the main human communicative means that are language and gesture. Thus, information presentation must be linked to natural language generation. Among natural language aspects, we want to emphasize the pragmatic aspects, and, in particular, the pragmatic forces (locutionary, illocutionary and perlocutionary forces, see [20] and [21]) that are conveyed together with a message. Since multimodality includes natural language, pragmatic forces will apply on each multimodal message and multimedia presentation act. In this subsection we show how illocutionary and perlocutionary forces can be handled by IMMPS.

3.3.1 Illocutionary force

When interpreting as well as generating, the message content is associated with an illocutionary force that expresses the act that is realized by the enunciation, and that depends on an underlying intention. Following Relevance Theory [22], 'saying that', 'telling to', and 'asking' are the main illocutionary forces. By 'saying that', the speaker expresses an **assertion** in order to make the addressee know something. By 'telling to', the speaker expresses a **demand** in order to make the addressee do something. By 'asking', the speaker expresses a **question** in order to know something from the addressee, with two cases: the close question or 'asking if' whose answer is yes or no, and the open question or 'asking WH-' whose answer is an information.

Concerning multimedia presentation, the way of presenting, for instance an alert, depends on the illocutionary force. If the system just wants to **inform** ('saying that'), it can use a certain method of presentation that totally differs from the method used to **encourage to act** ('telling to'). Moreover, the dialogue system may need a confirmation of the message reception. Then we can distinguish a 'saying that without feedback' from a 'saying that with a mandatory feedback'. The second corresponds to the use of the classical 'OK' or 'OK/cancel' dialogue boxes. As one of the contributions of our approach, we propose to model the two previous points with composite speech acts. That is not very far from the concepts of active presentation and passive presentation from [1], but here we emphasize the pragmatics of communication as it is modeled elsewhere in linguistics and computational linguistics work. One important point concerning the request for a feedback from the user is that such a behavior may be decided by the IMMPS with the help of the task manager. In fact, for some particular actions, a feedback request can be included in the task model. In such a case, IMMPS must not add an additional feedback request.

A distinction can be made between an explicit order and an implicit order. The 'OK' dialogue box constitutes an explicit order because it materializes the need of the system to get a confirmation of the reception of its message. On the other side, a message, for instance an alert, which aims at strongly encouraging action with no materialization of this goal constitutes an implicit order:

- Alert = "**saying that** problem" + "**telling to** react to it" (implicit order),

- Saying that with 'OK' feedback = "**saying that** information" + "**telling to** feedback" ('OK' explicit order),
- Saying that with 'OK' or 'cancel' feedback= "**saying that** information" + "**asking if** agree".

3.3.2 Perlocutionary force

Each message also has the aim of producing an effect on its addressee, whatever this effect is (just taking the message content into account, or realizing something precise). We claim that it is the dialogue manager that must manage the perlocutionary force, in particular the expected reaction following a demand from itself (next state in the user task model). It is the IMMPS that must correctly convey the perlocutionary aim, for instance by making a waiting attitude from itself obvious. As an example, an alarm that follows the detection of an inconsistency in the application database can have two aims: informing the user, i.e., something like "be careful, there is an inconsistency", or encouraging the user to give an information he is susceptible to know but has not yet passed on. In this case, a solution consists of opening a text box window, as an equivalent of an 'asking WH-' speech act. If the interface integrates an animated conversational agent, another solution consists of displaying an attitude that clearly conveys an expectation of the user's behavior.

In the case of a GUI, the perlocutionary force is linked to the graphical metaphors. In fact, the choice of the elements of the GUI has an influence on the user's future actions. As very simple examples, we are used to pushing buttons, and we try to write text inside each element that looks like a text box. In particular, when a table is displayed, we try to modify the content of the cells. In a general manner, we know that each displayed element has a function, and if we do not know that function we try to identify it. Consequently, the IMMPS must know the functions of all the GUI elements that it may have to present. Moreover, it must take these functions into account during the various phases of the presentation. For each GUI element, it must be aware of the input interaction possibilities. Then, it must inform the input events manager, and indeed the fusion module. To continue with the previous example, a table of numeric values can be presented using several methods depending on whether the values can be modified or not. First, the cells can be presented with a particular color or rendering, for instance with a grey tint if they are not modifiable. Second, each cell can be accompanied with a text box that makes the possibility of modification obvious.

Two additional aspects can be discussed on how the perlocutionary forces can be materialized considering the particularities of vocal interaction and natural language, for instance with mentional expressions. Considering the difficulties of speech recognition, several recognition grammars can be specified depending on the type of expected input utterances right after a multimedia presentation. Typically, a very general grammar, which by consequence is not very precise, is used when the system has to detect the theme in order to launch the related application. On the other hand, a command grammar is used when the user has the dialogue initiative. Another grammar that is specific to numbers, dates, and numeric values, is also used when the user must answer a question from the system that deals with such data. Consequently, IMMPS must

take into account the type of vocal feedback that is susceptible to follow a presentation act, and must inform the recognition module. For instance, the command grammar may be activated right after an inform-like presentation, and a specific grammar after a question-like presentation.

Concerning the particularities of natural language, we claim that the method to present multimedia information has an influence on the user's future linguistic choices. In particular, displaying pieces of information that follow an obvious order (arrival order or visual organization order) will favor **mentional expressions** such as "the first", "the second", "the next one", or "the last one". Likewise, displaying pieces of information that are obviously dissociated will favor **quantified expressions** such as "each", "all the". The consequences for the understanding of a linguistic message that follows a multimedia presentation are multiple. IMMPS must be aware that it may have to make obvious an ordering that was not expressed by the dialogue manager, for instance the default occidental vision order, from left to right and from up to down. IMMPS must also be aware of the way pieces of information are stuck together or not. Moreover, in such cases, IMMPS could inform not only the recognition module but also the module dedicated to natural language understanding.

3.3.3 Statement on communicative acts for IMMPS

With the four intentions that are (a) inform without feedback, (b) inform with mandatory feedback, (c) encourage to react, and (d) question, we propose the corresponding presentation acts classification:

- 'Inform': equivalent to the 'saying that' speech act, with no feedback required,
- 'Feedback inform': equivalent to the 'saying that' + 'telling to'/'asking if' composite speech act,
- 'Demand': equivalent to the 'telling to' speech act,
- 'Question': equivalent to the corresponding speech act, with the distinction between open question and closed question.

As an example of simple processing rules for IMMPS, the presentation act could depend firstly on the level of criticality: 'feedback inform' for a high level and 'inform' for a lower level. The act could also depend on the level of urgency: 'demand' for a high level and 'inform' for a lower level. This shows how pragmatics is related to information nature and useful to multimedia information presentation.

4 A TWO-STEPS PROCESS FOR IMMPS

4.1 First step: information repartition over the communication channels

The generation of output multimodal messages as well as the presentation of multimedia information is a process that aims at repartitioning the content of the message

or information over the communication channels, and at valorizing all partial contents within each communication channel. The main strategy for information repartition consists of taking into account a set of constraints and a set of preferences. Some additional strategies can then be imagined to optimize the generation of the message and favor its perception by the user.

4.1.1 Repartition by exploiting constraints

Constraints must of course be taken into account at the earliest phase of the presentation process. Some constraints are inherent to the information. As a typical example, the presentation of a map must be done graphically and not vocally (with the exception of the description of an itinerary over the phone, but in such a case the information is not really a map but an annotated description of a particular aspect of a map). Some other constraints are linked to the terminal, for instance when the terminal cannot produce vocal messages. There are also constraints that are associated with the presentation environment. In fact, environmental conditions can impose or forbid some of the communication channels. For instance, a strong ambient noise will forbid the vocal modality. Finally other constraints are linked to the user's abilities and roles. Concerning abilities, this is the case for some kinds of handicap. Concerning roles, it depends on the application and usual user profiles.

4.1.2 Repartition by exploiting preferences

Once the constraints have been taken into account, a lot of choices are possible and the IMMPS has to compute the various aspects of the interaction to make the most relevant choice. Among these aspects are: (a) the message content, with the exploitation of the communication channel that best fits the information constitution, and the preferential exploitation of several channels when the information is very complex, (b) the communicative act, with the preference of one single channel for a simple act and two channels for a composite act, (c) the interaction history, with a preference for the exploitation of the channel that has already been successfully exploited, (d) the user's preferences, which of course should be satisfied (the user can for instance prefer auditory feedbacks to visual ones), and (e) human factors. As an example of this last important aspect, displaying large information should be spread over time, in particular if reading it requires an important amount of persistent attention.

To compute the constraints and preferences and to choose the best information repartition paradigm, a lot of rule systems have been defined in the literature. Our purpose here is not to define another rule system, but to provide pragmatic and human factors related recommendations to design more natural IMMPS. These recommendations will have to be completed considering the design context. In particular, application specificities such as the data manipulated and the objectives of the user, will have to be carefully studied before defining the rule system. The rules will also be defined with preliminary consultations of ergonomics experts. As a basic example dealing with the characteristics of the message, here is a set of rules that can be considered as a basis:

- If the urgency of the message is fatal, then use both visual and auditory communication channels;

- If the urgency of the message is critical, then use the auditory communication channel as a priority;
- If the urgency of the message is nominal, then use either visual and/or auditory communication channel;
- If the information level of criticality is fatal, then use the auditory communication channel;
- Etc.

Another example of a rule system deals with the user's activity during the interaction:

- If the user is distant from the different interaction devices, then use both visual and auditory communication channels;
- If the user is close to the different interaction devices, then use either visual and/or auditory communication channels;
- If the user is attentive to the ongoing interaction, then use either visual and/or auditory communication channels;
- If the user is absent-minded, then do not use an auditory communication channel.

4.1.3 Making the link between distributed information

When a part of the information is displayed and another part is uttered, the user may not be able to make the link between the two parts and consider the information as a whole. But such a link is important because the two parts of the message must not be considered as two distinct messages, i.e., messages that can be treated separately. In their rule system, [24] emphasize this point with a rule for checking the presence of a semantic link between the visual part and the vocal part of the message. To the contrary of a VU meter or a video where the temporal synchronization is sufficient for the user to put together the sound track and the visual track, the association of a vocal message to a visual feedback might be seen as two distinct messages instead of one. Then, in this case there is the need for the system to provide some indications to the user so that he puts together the visual feedback (e.g., emphasizing a particular visual object) to the related part of the vocal message (e.g., a referring expression such as "this object").

A way to do this is to indicate with a modality that another part of the message is conveyed using another modality. An additional visual feedback can thus make the presence of the vocal message obvious. Using natural language, vocal messages such as "on the currently displayed map, you can see..." or "flight 102 is the one that flashes" include a reference to the visual modality. In fact, such a reference is a kind of 'deixis' and has been well studied in linguistics and computational linguistics works [14]. To us, an IMMPS has to handle, with particular care, deictic cases, in order to well manage the interactions between natural language and visual perception. As an example, the generation of "flight 102 is the one that flashes" can be seen as the following suite of processes:

1. The dialogue manager produces a presentation request using a logical form that corresponds to something like “make-obvious-to-the-user(flight-102)”;

2. The IMMPS chooses both a visual and vocal realization with the generation of a deixis, so that the user brings the two realizations together;

3. The IMMPS asks the natural language generation module to materialize the inter-modal deixis, i.e., the IMMPS indicates the nature of the display;

4. The natural language generation module produces the expression “the one that flashes”;

5. The IMMPS produces “Flight 102 is the one that flashes” and activates the visual flashing rendering.

Since the nature of the visual feedback is clearly explicit, this way of proceeding corresponds to an explicit inter-modal deixis, as opposed to the following implicit inter-modal deixis (using the same example but with another choice from the natural language generation module):

4. The generation module produces the expression “here is”;

5. The IMMPS produces “Here is the flight 102” and activates the visual flashing rendering.

Thus, deixis management is one important aspect of multimedia information presentation, and is integrated in our architecture of Figure 1.

4.1.4 Reinforcing the message by exploiting redundancy

When the level of urgency is high and in other cases of presentation, there is the need to reinforce the message in order to increase the probability of it being well perceived and assimilated by the user. This can be done by duplicating the information over two or all the communication channels, i.e., exploiting redundancy.

In fact, redundancy is the classical way to emphasize information when generating in a multimedia context. We want to soften this method, with the following arguments. First, there are of course a lot of arguments for the exploitation of redundancy: (a) if a communication channel does not work well, the other one makes up for it, (b) the more information is emitted, the more chances the addressee has to receive it, (c) the more information is presented again, the more chances the addressee has to become imbued with it. Second, these arguments can be opposed to human factors preoccupations that work against redundancy: (a) too many messages do not encourage the addressee to maintain his persistent attention, (b) too many messages increase the processing time and therefore the expected reaction delay. As an illustration, remember the famous example of an air crash due to a bad interpretation of redundant information: “– Why didn’t you answer the control tower who indicated to you that your landing gear was not out? – Because I had a klaxon that was sounding in my ears! – That’s incredible! That signal precisely indicated to you that your landing gear was not out!” As a statement, we propose the following basic but important rules to handle redundancy:

- Exploit redundancy only if the addressee should be able to make the link between the various emissions of the same information, i.e., if he can notice that it is redundancy;
- Do not exploit redundancy in the same communication channel (e.g., sound and voice like in the air crash example);
- When the message is so urgent or important that it cannot be ignored, be careful that redundancy does not introduce any perturbation.

4.1.5 Information repartition and multimodal fission

It is now well stated in the literature that managing input multimodality corresponds to the 'multimodal fusion' problem, and that managing output multimodality corresponds to the 'multimodal fission' problem [27]. But it is very rare to find work dealing with several levels of fusion or fission, with the aim to make precise what these processes are. We want to propose a unified vision of multimodal fusion and fission that is linked to our semantics and pragmatics related approach.

Multimodal fission consists of splitting the information into several parts considering the presentation aims, means and context. Now, information can be split at different levels. At the signal level, the information, considering its nature, is sent to the correct communication channel. This is typically the case for a video, the sound track being sent to the auditory channel and the visual track to the visual channel. This is also the case for a linguistic utterance accompanied by one or more deictic gestures, such as "I am putting that there" with two gestures, one for "that" and one for "there". In this example, IMMPS must be aware of the duration of the speech synthesis in order to provide the gestures, e.g., visual feedbacks, at the right moments. Splitting and synchronizing at the signal level is then a kind of multimodal fission, and is strongly linked to the constraint-based repartition over the communication channels.

At a semantic level, the information content can be dissociated over several modalities in order to better manage its complexity and to simplify the resulting monomodal messages. One important example related to human factors consists of displaying the part of the information that requires an important amount of persistent attention, and of verbalizing the part whose only aim is to capture selective attention. Splitting at a semantic level appears as another kind of multimodal fission, which is linked to the preference-based repartition.

At a pragmatic level, the message illocutionary force can be dissociated over several modalities in order to simplify the illocutionary force of each resulting monomodal message. For instance, a message labeled with a 'feedback inform' presentation act can be split into two messages: a first one that verbalizes the 'inform' and a second one that requires the 'feedback' using a text box. To us, this is a third kind of multimodal fission, as important as the previous ones, although it has not been studied in the literature.

To show the relevance of such a classification into three levels of fission, it is interesting to bring together fission with fusion. In fact, multimodal fusion can also be done at three different levels. At the signal level, the coordination of the various signals into one composite multimodal signal is a first kind of fusion. At a semantic level, the fusion of the message contents is a second kind of multimodal fusion, which is

strongly linked to the resolution of references to objects [14]. At a pragmatic level, the fusion of illocutionary forces corresponds to the fusion of events and is also a kind of multimodal fusion. To conclude:

- At the signal level there is **multimodal coordination** for input signal processing and **multimedia coordination** for output processing;
- At a semantic level there is **content fusion** for input message processing and **content fission** for output message processing;
- At a pragmatic level there is **event fusion** for input event processing and **presentation act fission** for output event processing.

4.2 Second step: information valorization

Once the output multimodal message has been repartitioned over the communication channels, there is the need to optimize and to valorize each piece of information within each communication channel. We can distinguish a main strategy that consists of taking into account a set of constraints and a set of preferences, and some additional strategies dealing in particular with human factors.

First, constraints have to be taken into account, with (a) the constraints that are inherent to the information, for instance the numbers of lines and columns when displaying a table, (b) the constraints that are linked to the terminal, e.g., the screen size with the same example, and (c) the constraints that are linked to the presentation environment, for instance a threshold for the ambient noise.

Second, preferences can be taken into account using a set of rules relying on: (a) the message content, we can imagine for instance display rules related to data structure, (b) the communicative act, for instance favoring a strong intensity for a 'demand' act, (c) the user's preferences, for instance displaying with a font size of 16 if it is a preference, (d) human factors, for instance exploiting the color red for an alert (because red is perceived faster than other colors).

Moreover, IMMPS should be able to optimize the information content within a modality. For instance, information can be spread out to the limits of the terminal, with rules like the following one: when displaying a picture, take all the available space. IMMPS should also have to emphasize a content, to exploit a salience. In this way, one important aim will be to adjust the communicative structure for putting one element into salience. When an avatar is used as a conversational animated agent, IMMPS may have to render emotions on contents, with the exploitation of the prosody and if necessary the multiple possibilities of the animated character. Lastly, to manage the user's attention, IMMPS should take into account the distinction between selective attention (that is captured by a transient verbalization or display) and persistent attention (that requires a persistent or permanent display).

5 CONCLUSION AND FUTURE WORK

In this paper we have proposed a set of theoretical and operational principles for the design of intelligent multimedia presentation systems. These principles and the related

classifications integrate the preoccupations from work dealing with adaptation to the terminal, to the environment, and to the user. Our proposal is based on the Speech Act Theory, and in a general manner on pragmatic preoccupations. Human factors are taken into account, and the foundations for more human-oriented systems are drawn. The method we propose for information presentation relies on two main phases, the first one consisting of the repartition of information over the communication channels, and the second one consisting of the valorization of each piece of information within each communication channel.

The recommendations we provide for the design of intelligent presentation systems have to be confronted with the applicative and design contexts in order to be translated into rule systems. In this paper our aim was not to present such rule systems, but to determine the underlying main preoccupations and methods, that can be applied to every kind of human-machine interactive system and not to a particular task-oriented one. Since we have focused mainly on theoretical aspects, a future paper will focus on technical details and implementations, with the presentation of a particular applicative context (an interactive support for cooperative decision making in the domain of air traffic management, which was at the basis of some of our observations and experimentations) and the related rule systems.

References

[1] E. André and T. Rist. Multimedia presentations: The support of passive and active viewing. In *Proceedings of the AAAI Spring Symposium on Intelligent Multi-Media and Multi-Modal Systems*, pages 22–29, Stanford, 1994.

[2] Y. Arens and E. Hovy. How to describe what? towards a theory of modality utilization. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 487–494. Lawrence Erlbaum Associates, 1990.

[3] J.L. Beckham, G.D. Fabbrizio, and N. Klarlund. Towards SMIL as a foundation for multimodal, multimedia applications. In *Proceedings of EUROSPEECH 2001*, pages 1363–1366, Aalborg, Denmark, 2001.

[4] N.O. Bernsen. A reference model for output information in intelligent multimedia presentation systems. In *Proceedings of the ECAI'96 Workshop on: Towards a Standard Reference Model for Intelligent Multimedia Presentation Systems*, Budapest, Hungary, 1996.

[5] M. Bordegoni, G. Faconti, S. Feiner, M.T. Maybury, T. Rist, S. Ruggieri, P. Trahanias, and M. Wilson. A standard reference model for intelligent multimedia presentation systems. *Computer Standards and Interfaces*, 18, 1997.

[6] C. Cadoz. *Les réalités virtuelles*. Flammarion, Paris, 1994.

[7] W. Chou, D.A. Dahl, M. Johnston, R. Pieraccini, and D. Raggett. EMMA: Extensible multi-modal annotation markup language. Available at http://www.w3.org/TR/emma/, 2002.

[8] J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R.M. Young. Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. In *Proceedings of INTERACT'95*, Lillehammer, Norway, 1995.

[9] J.J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[10] P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Speech Acts, Syntax and Semantics (Vol. 3)*, pages 41–58. Academic Press, New York, 1975.

[11] J. Itten. *The Art of Colour*. Reinhold Publishing Corp., New York, 1961.

[12] C. Karagiannidis, A. Koumpis, and C. Stephanidis. Adaptation in intelligent multimedia presentation systems as a decision making process. *Computer Standards and Interfaces*, 18(6-7), 1997.

[13] W. Kohler. *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. Liveright Publishing Corp., New York, 1947.

[14] F. Landragin. Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems. *Signal Processing*, 86(12):3578–3595, 2006.

[15] I. Mel'čuk. *Communicative Organization in Natural Language: The Semantic-Communicative Structure of Sentences*. Benjamins, Amsterdam, 2001.

[16] G.A. Miller. The magical number seven, plus ou minor two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

[17] H. Prendinger, S. Descamps, and M. Ishizuka. MPML: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages and Computing*, 15(2):183–203, 2004.

[18] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.

[19] C Rousseau, Y. Bellik, and F. Vernier. Multimodal output specification/simulation platform. In *Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, pages 84–91, Trento, Italy, 2005.

[20] J.R. Searle. *Speech Acts*. Cambridge University Press, Cambridge, 1969.

[21] J.R. Searle. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge, 1979.

[22] D. Sperber and D. Wilson. *Relevance. Communication and Cognition*. Blackwell, Oxford, 2nd edition, 1995.

[23] R. Stevenson. The role of salience in the production of referring expressions: A psycholinguistic perspective. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 167–192. CSLI Publications, Stanford, 2002.

[24] A. Sutcliffe and P. Faraday. Designing presentation in multimedia interfaces. In *Proceedings of CHI'94, Conference on Human Factors in Computing Systems*, pages 92–98, Boston, 1994.

[25] W3C. Multimodal interaction activity, multimodal interaction working group. Available at http://www.w3.org/2002/mmi/.

[26] W3C. Synchronized multimedia integration language (SMIL) specifications, SMIL 2.1 proposed recommendation. Available at http://www.w3.org/AudioVideo/.

[27] W. Wahlster. Smartkom: Fusion and fission of speech, gestures, and facial expressions. In *Proceedings of the First International Workshop on Man-Machine Symbiotic Systems*, pages 213–225, Kyoto, Japan, 2002.