Dialogue History Modelling for Multimodal Human-Computer Interaction

Frédéric Landragin and **Laurent Romary**
LORIA – UMR 7503
Campus scientifique – B.P. 239
F-54506 Vandœuvre-lès-Nancy Cedex – France
{landragi,romary}@loria.fr

Abstract

The design of multimodal dialogue systems requires a particular attention on the way of managing dynamic heterogeneous information. We present a theoretical model of a multimodal dialogue history, that includes a global history and local histories linked to the various modalities. We describe the nature of these components and the relations they entertain. To save the information and to exploit and confront them during the interpretation of an utterance from the user, we need an unified representation format. We developed the MMIL (MultiModal Interface Language) model that we present here with two main aspects: the representation of a simple utterance and the representation of a dialogue structure. Then, we draw some conclusions concerning the exploitation of this framework in the OZONE system, with an interest on the management of attentional scores inside the dialogue history.

1 Introduction

A dialogue history is always needed in oral understanding systems, for example to resolve ellipses and anaphors. The interpretation of an utterance can exploit the previous utterances that are saved in this history. When the task includes objects that can be referred to, this history must keep the objects identity along with the expressed utterances. Problems can arise in multimodal systems, where gesture and speech are combined and linked to the visual perception of the displayed scene. What type of information has to be saved to resolve references to objects? Are there separate histories for visual perception, gesture, speech, referents, reference domains? How can a history take forgetting phenomena into account? In this paper, we propose some theoretical leads to these infrequently debated topics. The particular aspects we want to address are the following ones:

- Components of the dialogue history: what is a dialogue history and how is it shared considering the modalities in natural language and multimodal systems (visual perception history, linguistic history, etc.).
- Nature and roles of the global history: temporal marks and pointers to local histories, and, eventually, a model of attention and forgetting (i.e., the former information are forgotten step by step). First main problem: limits of histories. We cannot save all information that can help the reference resolution and the utterance interpretation (information structure, task marks, interaction marks, etc.). We need to save only the information that have been the object of a computation during a previous phase of the dialogue. Second main problem: a model of the user's attention may be useful. The idea is to identify a focused part of the histories, that corre-

sponds to the part which is at the moment in the mind of the user. Third main problem: we need a model of forgetting. Idea: when the user talks about a blue triangle, the concepts that are linked to "triangle" and to "blue" are activated. Their activation rates will then decrease step by step, as the time goes by (except if they are activated again, i.e., mentioned in the dialogue).

- Unified representation of information in the various histories: first through the systematic notion of reference domain (see Landragin & Romary, 2003), second by using a unified representation format (MMIL: MultiModal Interface Language) which we describe in this paper.
- Illustration and applications of the model we propose in the framework of the OZONE European project.

2 Components of the dialogue history

The main purpose of the linguistic history is to keep the trace of referring actions. The information to be saved are referring expressions, referents, and reference domains. Referring expressions have to be saved for a further exploitation of the referent accessibility. Referents have to be saved because of evolutive referents and objects deletions (see the interpretation of "it" in Example 1). Concerning reference domains, Example (2) shows the linguistic construction (using the coordinating conjunction "and") of a reference domain including two tables. This domain is kept in the history and constrains the resolution of a further referring expression. (2a) is then authorized, but not (2b). Another interest of reference domains is to propose a default set of objects for the interpretation of "other one" expressions. Reference domains are linked to each others, in order to model referential and anaphoric chains (Salmon-Alt, 2001).

(1) Remove the big desk. Replace it with a round table.
(2) Add a wooden table and a black plastic table.
(2a) Put the wooden table on the left and remove the other one.
(2b) *Put it on the left.

The visual history keeps the successive states of the scene. The information to be saved are the objects and their properties (including the coordinates), the perceptual groups, and their structuring (one group can include several groups). Visual salience and focus spaces (Beun & Cremers, 1998) are modelled into visual reference domains (Landragin, 2001). The visual history is necessary to face to situations such as (3). In this example, the interpretation of "underneath" needs a return to a previous state of the scene. As the pointing gesture was linked to the visual context, it was also kept in this history. Determining the position of the shelf uses a combination of the coordinates of the painting and of the gesture trajectory.

On the importance of the gestural part in the visuo-gestural history: when the user produces several times the same type of trajectory, it is more and more easy to interpret (even if they are more and more ambiguous, more and more imprecise, more and more quick, or less close to the target objects).

(3) Remove this painting (+ *gesture*). Add a shelf underneath.

The task history groups the performed actions, the referents to which they applied at the time, and links between these objects and the task's purpose and sub-purposes. Following (Grosz & Sidner, 1986), we take intentional structures into account, and we model them as task-linked reference domains. We show in a complex multimodal example (extracted from Ozkan corpus and described in details in the paper) how keeping these domains can be useful for the dialogue understanding.

3 Nature of the global history

One important point is that linguistic, visuo-gestural and task-linked reference domains are all structured in the same way. Consequently, domains can be confronted and integrated. That is a strong point of this model compared to heteroge-

neous modality-dependent theories. However, the global dialogue history does not correspond to the integration of all local histories, but includes the following elements: pointers to parts of them, results of the referring actions, *a posteriori* evaluations, system's reactions, and time stamps. Local histories can then be seen as specialized agents, and the global history as the coordinating agent.

The last problem deals with the storage capacity of the global history. With a cognitive concern, we can consider the history as a model of short-term memory, and limit its capacity to the seven most recent items, whatever the nature of these items (Miller, 1956). More recent works in psychology tend to limit this capacity to five or only four items (see Rousselet & Fabre-Thorpe, 2003). But since numerous items can be accessible, the working of the history appears to be much closer to long-term memory. The problem that arises is that all information has to be saved and is at the same level of accessibility. We prefer to consider the global history as a model of forgetting. The longer an entity has been the focus of user's attention, the more important it is to emphasize it in the history. We thus propose a methodology for tracking attentional scores, that apply to every object, category, event, or property:

- the more the user refers to an object, the greater the object's attentional score;
- the more he evokes a property (like a red colour), the greater the score of its related concept ("red");
- the more he performs an action, the greater the score of the corresponding software primitives (for example, a function or a class).

The score of an entity involves the scores of all its properties. These scores are managed so that the entities with the best scores are favoured during the interpretation process. Scores also (like human attention) decrease as the dialogue progresses. This approach has some common aspects with linguistic work like that of (Ariel, 1988) or (Lappin & Leass, 1994), but is more inspired by psycholinguistics. More precisely, we want to extend the principle of the Logogen Model (Morton, 1982). Our propositions concerning local and global histories seem to fit well such an aim. Though this is on-going research (no concrete algorithm has yet been implemented), the nature of information to be saved is already sufficiently precise to deduce the main characteristics of an algorithm and of a representation format.

4 Representing semantic content: MMIL

To represent the various histories, we need a unified representation format. The purpose is to confront in the global history information from different histories. This confrontation is only possible if the heterogeneous information are represented into similar structures. So we need a model to represent visual, gestural, linguistic and task-linked information in a same manner. For that, we use the MMIL model (MultiModal Interface Language) that was designed for the MIAMM European project and that was updated for the OZONE European project (see references).

The MMIL meta-model abstracts different levels of dialogue information (phone, word, phrase, utterance) by means of a flat ontology, which identifies shared concepts and constraints. The definition layer of the ontology includes two kinds of entities: events and participants. Events are objects associated to the temporal level, while participants are static entities acting upon or being affected by the events. Dependencies between entities are represented as typed relations linking structural nodes. Contrary to other semantic information models, the MMIL meta-model does not include relations, which are perceived as qualifying descriptors defining anchors among entities. As the other information units of the MMIL model (e.g., morpho-syntactic, domain, annotation descriptors), relations act in the information architecture as a set of descriptors (data categories) that formally describe the specification constraints. The data categories, expressed in an RDF format compatible with ISO 11179-3, give the necessary openness to the design of the semantic structures, so to cope with the potential flexibility of the model.

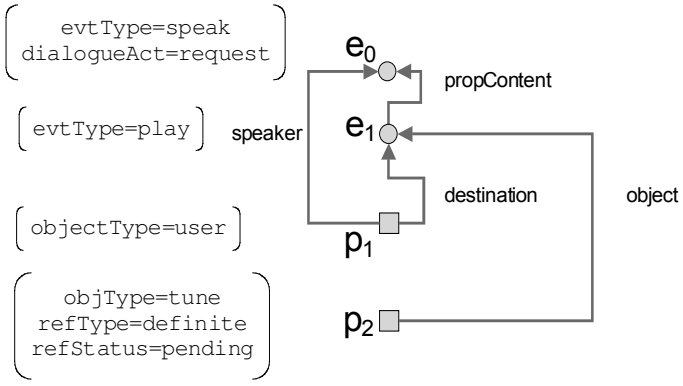
4.1 Simple utterance representation

One of the central purposes of MMIL is of course to be able to represent the actual semantic content of the various utterances processed by the linguistic modules in the MIAMM and Ozone architectures. To do so, we framed a core organization for MMIL structures at the output of linguistic modules that articulates:

- A specific event e0 that systematically represents the speech event associated with the utterance. Being categorized as such (/event type/=/speech event/), the event is anchored on temporal node that informs its beginning and ending date, and it may be further refined by various data categories corresponding to the /speaker/, /addressee/ and /speech act/;
- The actual (possibly underspecified) meaning of the utterance represented by an event e1 corresponding to the main predicate expressed by the utterance and which is related to e0 by a /propositional content/ relation. The event e1 is in turn associated with all the necessary descriptive elements such as its actual arguments, which are represented by one or more participants associated to it by the basic semantic roles (/agent/, /patient/, /location/, etc.) identified by the linguistic parser.

The fact that a full representation is provided for the speech event instead of just providing the corresponding propositional content offers several advantages that by far compensate the little extra complexity that it brings to the representation. First, it is an essential aspect in dialogue management (see below), to be able to relate an utterance to an other, and it may not be possible to make this boil down to the sole organization of contents. Second, it provides a clear and coherent background for personal and temporal deixis interpretation, which can be directly related to the information available at speech event level, rather taking up information maintained specifically to this purpose. Finally, it is an essential basis unifying references to the application domain and to the discourse proper. An utterance such as "Please repeat" can only be processed if a homogeneous treatment is made of speech event within the space of the various events expressed along the course of a dialogue.

As an example, Figure 1 shows a graphics summarizing the main component of the utterance "Play me the song", together with the corresponding full XML representation.




```
<mmilComponent>
  <event id="e0">
    <evtType>speak</evtType>
    <dialogueAct>request</dialogueAct>
    <speaker target="p1"/>
  </event>
  <event id="e1">
    <evtType>play</evtType>
  </event>
  <participant id= "p1">
    <objType>user</objType>
  </participant>
  <participant id= "p2">
    <objType>tune</objType>
    <refType>definite</refType>
    <refStatus>pending</refStatus>
  </participant>
  <relation source="e1" target="e0"
            type="propContent"/>
  <relation source=" p1" target="e1"
            type="destination"/>
  <relation source= "p2" target="e1"
            type="object"/>
</mmilComponent>
```

Figure 1. Graphical representation of the MMIL structure associated to the utterance "Play me the song!".

Furthermore, the inherent hierarchical structure of events and participants in the MMIL metamodel, together with the actual reification of the speech event in any language related MMIL representation, allows the representation of more complex phenomena related to the actual segmentation of the spoken input. Indeed, it may occur that what is considered as one single input at speech acquisition level may well be further segmented at speech recognition or parsing level

as several sub-utterance bearing, for instance, specific dialogue acts or propositional content. In those cases, the speech event is further subdivided into the necessary components, as exemplified in the following simplified representation for "No, to Paris" (Figure 2).

```
<mmilComponent>
  <event id="e0">
    <evtType>speak</evtType>
    <speaker target="p1"/>
    <event id="e0-1">
      <dialogueAct>reject</dialogueAct>
    </event>
    <event id="e0-2">
      <dialogueAct>inform</dialogueAct>
    </event>
  </event>
  ...
</mmilComponent>
```

Figure 2. Representation for "No, to Paris."

As can be seen, both the information related to the event type and speaker are factored out at the main event level, whereas the dialogue act information is, in this case, specifically attached to the sub-components of the utterance.

4.2 Dialogue structure in MMIL

As in any classical man-machine dialogue architecture, the dialogue manager is in charge of both contextualizing each utterance coming from lower level linguistic modules (e.g. by interpreting referring expressions) and maintaining the overall logics of the dialogue, through, among other thing, a proper management of the dialogue structure. As a matter of fact, dialogue structure is the result of combining sentence level information with higher principles of discourse organization, which also closely interacts with the actual semantic content of utterances, among which focusing information take some specific importance. Besides, dialogue management involves being able to put in relation the user's utterances with the actual decisions or actions (spoken feedback, information presentation, actions at application levels) taken by the dialogue manager itself (or its action planning component). This is again where the homogeneous background provided by MMIL offers a flexible way of dealing with those various phenomena, under the condition that some clear principles are stated to maintain the coherence across the various information sources, that is at user's input level, dialogue internal processing level, and discourse management level.

At user's input level, we have already mentioned that each utterance is initially represented by a speech event to which is attached the actual semantic content as understood at the parsing stage. The speech event is qualified by at most one dialogue act taken from a basic typology of six values, i.e. /opening/, /closing/, /inform/, /query/, /accept/, /reject/. Those values are considered as surface dialogue acts, as they result from the sole analysis of the utterance inner structure prior to any specific contextualization of the semantic content. They correspond to the core (and consensual…) values that can be observed from various systems or annotation schemes (e.g. DAMSL, HCRC, etc.) that have been around in the last decades. They also correspond (except for the obvious /opening/ and /closing/) to the basic dialogues acts identified in the work by FIPA on inter-agent communication, which, as we shall see, contributes to a more homogeneous treatment of events in our dialogue architecture. Appendix A provides the ISO 11179 conformant description of those 6 dialogue acts in the perspective of stabilizing those values within the man-machine dialogue community.

As a consequence, we can also represent each event occurring within the dialogue architecture proper as a MMIL event, which bears similar characteristics in common with users' speech events. Dialogue internal event are thus represented in MMIL by means of several core characteristics:

- They are typed according to a basic ontology of dialogue management action combining general purpose actions (e.g. spoken feedback to the user, graphical presentation of information) and application specific primitives (e.g. queries to an underlying database);
- Like any other MMIL event, they can contain a temporal anchor indicating either when the action has taken place or when it is to be taking place (in the case it is still in a pending state within the dialogue architecture);

- They are qualified by one of the following four dialogue acts: /inform/, /query/, /reject/, /accept/, the last two are acts being used to validate or invalidate (e.g. when the information is not available or a service is down) an initial inform or query from one module to another. It should be noticed here that in both the MIAMM and OZONE architecture, a clear difference is made between the technical management of communications flows between modules, as can be typically handled by SOAP mechanisms in the context of a web service based technological deployment, with the management of the exchanges between modules from a semantic point of view. For instance, the same kind of behavior is not to be expected when a module within the architecture is physically down and when it has the knowledge cannot deliver a certain service that has been asked to it;

- Such dialogue internal event can be generalized to be used to communicate to external processes that may provide services in relation to the dialogue underlying task.

The example of Figure 3 shows the simple representation of the master event associated to a query issued by the action planner to the MIAMM database.

```
<event id="e4">
  <evtType>database query</evtType>
  <dialogueAct>query</dialogueAct>
  <evtStatus>actuated</evtStatus>
  <tempSpan startPoint="2004-04
     -05T17:00:00" endPoint="2004-04
     -05T17:00:01"/>
</event>
```

Figure 3. Representation of a master event.

Finally, MMIL structures can be used, on the basis of what has thus been presented, to uniformly represent dialogue structures that have been construed at dialogue management level. Without entering into the details of the supporting arguments for doing this, it seems by far more appropriate to base dialogue structure representation on inter-event relations (discourse relations, when dealing with users' input) then try to infer deep dialogue acts from the user's utterances. In this context, each time the dialogue manager infers a connection between any two events in the course of dialogue, it can report about it (for instance to the action planner module) by means of simple MMIL structures combining those events. As an example, a basic acknowledgement by the user ("Fine for me") to a proposal by the system ("I have this song from the Beatles") will be reported by a MMIL structure such as follows (Figure 4).

```
<mmilComponent>
  <event id="e1">
    <evtType>speak</evtType>
    <speaker target="p2"/><!-- system-->
    <dialogueAct>inform</dialogueAct>
  </event>
  <event id="e2">
    <evtType>speak</evtType>
    <speaker target="p2"/><!-- system-->
    <dialogueAct>accept</dialogueAct>
  </event>
    ...
  <relation source="e2" target="e1"
     type="confirm"/>
</mmilComponent>
```

Figure 4. Representation of a link between two events.

4.3 Standardization in the domain of semantic content representation

The work we have conducted on the definition of the MMIL language can be seen as a kind of experiment to identify precise requirements on what a general framework for multimodal content representation. Those requirements should obviously go beyond what has been described in (Bunt & Romary, 2002), in order to identify classes of applications which bear enough features to be covered by one single model. Indeed, it may not be likely that the kind of representations needed for such applications as information extraction, named entity recognition, reference annotation, or the annotation of temporal structure will be based on exactly the same underlying structures. Still, it seems necessary that those various types of models do share a common semantics for any sub-structure they would share and even more for any elementary descriptor they would use (e.g., a certain dialogue act /inform/, or discourse relation /elaboration/, a temporal relation /overlap/, or an elementary role in relation to an event /agent/). Such a goal obviously requires that there is some kind of consensus on providing some shared definition of such concepts, as well as an international infrastructure to submit, select and

disseminate those descriptors. The first aspect is one of the topics which has been considered as underlying the activity of the ACL/SIGSEM working group on multimodal semantic content representation and is being pursued through a series of meetings that have taken place since November 2002.

The second aspect is the core of a standardizing effort in ISO committee TC 37 to deploy an online data category registry intended to cover a wide variety of descriptors (also known as *data categories* in the TC 37 terminology) identified in existing representation or annotation practices. In this context, we would like to see MMIL as one instance of such a descriptive and modelling activity which would nicely fit the needs of multimodal dialogue system when conveying meaning from one component to another, and when managing meaning inside a component (and particularly inside the dialogue history). If it is the case, we could also contemplate using MMIL—or a dialect thereof—for such tasks as the evaluation of dialogue systems.

5 Applications of the model

A first application of the attentional scores that we propose is a help during the resolution of verbal ellipses. For instance, in a man-machine dialogue system that consists in the interrogation of a music database (queries about authors and songs, and commands like "play my favourites", see MIAMM European project for the implementation of such a system), when the user asks several times to "play this tune", the attentional scores of "play" and of "tune" are maximal. Then, if the user produces the incomplete utterance "this tune", the system can easily infer that the requested action is "play". This method for exploiting attentional scores can also be used for the resolution of anaphoric expressions, like "play this one" after "play this tune".

A second application is to allow more spontaneous reactions of the dialogue system, by using recalls when some information may have been forgotten. For instance, in reservation tasks like the reservation of a train ticket or a room in a hotel (see OZONE European project), one of the purpose of the dialogue is to specify a number a parameters that allow to launch a reservation request. For a train ticket like in our OZONE's demonstrator, the parameters are the departure station, the destination, the way (including changes of train), and the time (of departure or arrival). For a room reservation, the parameters are the number of persons, the date of arrival, the date of departure, and some options (breakfast, etc.). In OZONE, the user can begin the dialogue with "I want to go to Paris", and then can ask for some information about the possible ways, their duration, the changes, etc. The resulting dialogue can be quite long, and the destination (Paris) that has only been mentioned once at the beginning can have a very low attentional score. Thus, the system may be aware of this low score and may produce a sentence like "Do you still want to go to Paris?" or "You confirm a train ticket to Paris, don't you?". This is important, not only because it reactivates the salience of the destination, but also because it adds a collaborative aspect to the dialogue. Even if the system has not forgotten the destination, it's important for it to show a human-like cognitive behaviour. Of course, experimentations have still to be done to determine if subjects who interact with such a system feel (or not) a kind of strangeness in the reactions of the system.

6 Conclusion and future work

For now, our participation to the IST-OZONE European project has consisted in the realization of a dialogue system demonstrator for a transport information service task. The research work we have described in this paper will allow us to improve this demonstrator and to provide a second system. For this improvement, we will focus (among other points) on the design of the dialogue history using attentional scores. The OZONE's application appears to be an efficient framework for that. As other future works, we want to test our model in generation systems, and for other modalities like written texts. To conclude, we want to show with this experience that designing multimodal systems with spontaneous communication abilities has to make the most of linguistic and psychological results, and that a crucial point in this design is the representation of information that is managed by the system during the dialogue.

References

Mira Ariel. 1988. Referring and Accessibility. *Journal of Linguistics*, 24:65-87.

Robbert-Jan Beun and Anita Cremers. 1998. Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition*, 6(1/2):121-152.

Harry Bunt and Laurent Romary. 2002. Towards Multimodal Content Representation. In K. Lee, K. and K. Choi (Eds.) *Proceedings of LREC 2002 Workshop on International Standards of Terminology and Linguistic Resources Management*, pp. 54-60.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.

Frédéric Landragin. 2001. Visual Salience and Perceptual Grouping in Multimodal Interactivity. In: *First CLASS Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 151-155.

Frédéric Landragin and Laurent Romary. 2003. Referring to Objects Through Sub-Contexts in Multimodal Human-Computer Interaction. In: *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck'03)*, Wallerfangen, pp. 67-74.

Shalom Lappin and Herbert J. Leass. 1994. A Syntactically Based Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535-561.

MIAMM, Multidimensional Information Access using Multiple Modalities. IST-2000-29487 European Project. Website: http://www.miamm.org/.

John Morton. 1982. Disintegrating the Lexicon: An Information Processing Approach. In: J. Mehler, E. C. T. Walker and M. F. Garrett (Eds.) *On Mental Representation*. Erlbaum, Hillsdale, NJ, pp. 89-109.

OZONE (O_3), Offering an Open and Optimal roadmap towards consumer oriented ambient intelligence, IST-2000-30026 European Project. Website: http://www.extra.research.philips.com/euprojects/ozone/.

Guillaume A. Rousselet and Michèle Fabre-Thorpe. 2003. Les mécanismes de l'attention visuelle. *Psychologie Française*, 48(1):29-44.

Susanne Salmon-Alt. 2001. Reference Resolution within the Framework of Cognitive Grammar. In: *Proceedings of the Seventh International Colloquium on Cognitive Science* (ICCS'01), San Sebastián, Spain.

Appendix A: MMIL core dialogue acts

This annex describes the possible values of the **/dialogue Act/** data category. It is a selection of the dialogue act listed in the literature, which appears to us as sufficient for interpreting users' utterances in MIAMM (as opposed to annotation tasks, which would have probably required a more elaborate scheme).

Dialogue acts are usually described within an extensive hierarchy providing comprehensive groupings for them. It is not the intention of this simple typology to describe such a hierarchy, even if we have tried to organize thee list on the basis of some general dialogical categories. Further work within ISO/TC37/SC4 should incorporate this aspect more neatly.

Discourse management/Conventional acts:

/Opening/ **Def:** An utterance or segment establishing the communicative contact between a speaker and an addressee. **Note:** also know as 'Greet.'

/Closing/ **Def:** An utterance or segment finishing the communicative contact between a speaker and an addressee.

Initiative:

/inform/ **Def:** The speaker provides information to the user. **Note:** known under various names depending on the encoding scheme; Update (LINLIN), Explain (HCRC) Statement (DAMSL), Inform (TRAINS).

/request/ **Def:** The speaker aims to get the addressee to perform some action. **Note:** know under various names depending on the encoding scheme; Instruct (HCRC), Influencing Addressee Future Action (DAMSL), Action-directive (DAMSL), Request (TRAINS), Open-option (DAMSL), Suggest (TRAINS). No distinction is made here between Request and Suggest. Queries are represented as requests.

Response/Backward Looking Function:

/accept/ **Def:** The speaker agrees to all of the antecedent. **Note:** Corresponds to 'confirm_positive'. No distinction is made here between 'accept' and 'accept-part' (as in DAMSL).

/reject/ **Def:** The speaker disagrees with all of the antecedent. **Note:** Corresponds to 'confirm_negative'. No distinction is made here between 'reject' and 'reject-part' (as in DAMSL).