Types of cooperation and referenceable objects implications on annotation schemas for multimodal language resources

Jean-Claude Martin*

* LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
martin@limsi.fr

Abstract

The aim of this workshop is to discuss coding schemes to add annotation to multimodal language resources (MLR) with a view to developing an international standard. In this paper, we present some of the requirements on such coding schemes. We believe that they should enable users to search for the six types of cooperations between modalities that we propose as a typology for MLR. We explain how such "high-level" primitives could be used for testing the power of a standard for MLR annotation.

1. A framework for analysing multimodal behavior

1.1. Multimodal Wizard of Oz experiments

In order to reach usable multimodal human computer interfaces (i.e. combining several communication modalities such as speech and gesture), knowledge about the multimodal behavior of future potential users is needed.

In this perspective, several researchers around the world are recording users interacting with multimodal prototypes or simulated systems. Yet, there is still a need for multimodal metrics enabling the behavioral analysis of such corpora.

Multimodal Wizard of Oz experiments, such as the one made by the Stanford Research Institute (Kehler et al. 98), are providing observation corpora which are more complex to analyse than the simple interactions enabled in existing multimodal systems (table 1).

For a review of several multimodal Wizard of Oz experiments, see (Martin et al. 98). The existing metrics that we have found in the literature for analysing users behaviour do not try to quantify the continuum between redundant and complementary use of speech and gesture.

1.2. The TYCOON typology

In a first step, we have proposed to use the TYCOON typology that we had initially developped for studying multimodal systems (Martin & Béroule 93). According to this typology, several modalities may cooperate by: equivalence, transfer, specialisation, redundancy, complementarity, or concurrency.

1.2.1. Equivalence

A cooperation by equivalence is defined by a set of modalities, a set of chunks of information, which can be produced by either of the modalities and a criterion, which is used to select one of the modalities. When several modalities cooperate by equivalence, this means that a chunk of information may be produced as an alternative, by either of them.

1.2.2. Transfer

A cooperation by transfer is defined by two modalities and a function mapping the output of the first modality into the input of the second modality. When several modalities cooperate by transfer, this means that a chunk of information produced by one modality is used by another modality.

1.2.3. Specialization

A cooperation by specialization is defined by an modality, a set of modalities *A* and a set of chunks of information this modality is specialized in when compared to the modalities of the set *A*. When modalities cooperate by specialization, this means that a specific kind of information is always produced by the same modality.

1.2.4. Redundancy

Several modalities, a set of chunks of information and three functions define a cooperation by redundancy. The first function checks that there are some common attributes in chunks produced by the modalities, the second function computes a new chunk out of them, and the third function is used as a fusion criterion. If modalities cooperate by redundancy, this means that these modalities produce the same information.

1.2.5. Complementarity

A cooperation by complementarity is defined similarly as a cooperation by redundancy except that there are several non-common attributes between the chunks produced by the two modalities. The common value of some attributes might be used to drive the fusion process. When modalities cooperate by complementarity, different chunks of information are produced by each modality and have to be merged.

1.2.6. Concurrency

A cooperation by concurrency means that several modalities produce independent chunks of information at the same time. These chunks must not be merged.

1.2.7. TYCOON's limitations

When trying to apply our typology for analysing the multimodal behavior of subjects, we had difficulties for providing a fine grain analysis of most multimodal observations which can be neither qualified as redundant

or complementary, but are indeed in between these two extremes.

1.3. Referenceable objects and salience values

In order to enhance our typology, we have introduced the notion of "referenceable object". A referenceable object is an objet of a graphical application (i.e. an hotel icon in a map application) wrapped with multimodal knowledge : static knowledge (linguistic or gestural clues on how to refer to this object) and dynamic knowledge (salience values used for solving references).

The salience value of a referenceable object in one modality is proportional to the importance according to which this object seems to be referred to in this modality in a given utterance. Salience values have already been used in multimodal interfaces by (Huls et al. 95), but without referenceable objects and types of cooperation as we do. We have proposed that some rules for the computation of these salience values should be defined for each modality (table 2). A software prototype using such rules for soolving references between spoken utterances and 2D simple gestures has been implemented (Martin and Néel 1998).

1.4. Metrics for measuring multimodal behavior

These rules are involved in the computation of metrics measuring the rate at which the user makes use of the different types of cooperation between modalities.

The rate at which a subject makes use of equivalence (i.e. switches between several modalities for the same command) is computed with the following formula: the number of commands C_j expressed via different modalities is divided by the total number of commands expressed by the subject during the experiment.

$$\tau_{equivalence} = \frac{\left|\{Cj / equivalent(Cj)\}\right|}{\sum_{Cj} \left|Cj\right|}$$

The rate at which the subject's behavior is either redundant or complementary is computed with the following formula: a global salience value is computed over all referents r_k of all the commands C_j expressed by the subject; then this number is divided by the number of referents expressed by the subject during the experiment.

$$\tau_{compl./\,redund.} = \frac{\sum_{Cj} \sum_{rk \in R(Cj)} salience(Cj, rk)}{\sum_{Cj} \left|R(Cj)\right|}$$

Some results of the computation of such statistics are described in (Kehler et al. 98).

2. Implications on annotation schemas for MLR

As more and more multimedia and multimodal information becomes available through computers, annotation schemas are of importance in order to enable searching.

The Linguistic Annotation page[1] lists a number of schemes for linguistic or gesture annotation.

We think that in order to build a standard for the annotation of multimodal ressources, one needs to make a in depth study of what is multimodality, not only to build a mixture of existing schemes.

What we suggest is that the future standard for annotation schemas of multimodal resources should:

- enable searching for examples of cooperation between modalities (i.e. finding out where audio and graphics cooperate by equivalence and why),
- include annotation about referenceable objects (i.e. how many objects around people using speech and gestures).

3. References

Huls, C., Claassen, W., Bos, E., 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*. Volume 21, Issue 1, March 1995. Pages 59-79.

Kehler, A., Martin, J.C., Cheyer, A., Julia, L., Hobbs, J., Bear, J., 1998. On Representing Salience and Reference in Multimodal Human-Computer Interaction. *Proceedings of the AAAI'98 workshop on Representations for Multi-modal Human-Computer Interaction*. July 26-27, 1998, Madison, Wisconsin. USA http://tigger.cs.uwm.edu/~syali/AAAI-98-Workshop/aaai-wrkshp.html

Martin, J.C., D. Beroule., 1993. Types et Buts de Coopération entre Modalités. *Actes des cinquiemes journées sur l'ingénierie des Interfaces Homme-Machine (IHM'93)*, 19-20 octobre, Lyon, France

Martin, J.C., Julia, L. & Cheyer, A., 1998. A Theoretical Framework for Multimodal User Studies Proceedings of the Second International Conference on Cooperative Multimodal Communication, Theory and Applications (CMC'98), 28-30 January 1998, Tilburg, The Netherlands http://www.limsi.fr/Individu/martin/publications/download/cmc98-2.ps

Martin, J.C. and Néel, F. 1998. Speech and gesture interaction for graphical representations: theoretical and software issues. In Proceedings of the Workshop on "Combining AI and Graphics for the Interface of the Future", European Conference on Artificial Intelligence (ECAI'98). August 24, Brighton.

Martin, J.C. 1999. TYCOON:six primitive types of cooperation for observing, evaluating and specifying cooperations *Working notes of the AAAI Fall 1999 Symposium on Psychological Models of Communication in Collaborative Systems* November 5-

[1] http://morph.ldc.upenn.edu/annotation

7th, 1999, Sea Crest Conference Center on Cape Cod, North Falmouth, Massachusetts, USA. http://www-sop.inria.fr/acacia/PM/

Graphics	Ontario Place selected. Several other sites displayed.
Speech	How far is Ontario Place from my hotel ?
Drawing	Circle Ontario Place during the utterance of *Ontario Place*
Writing	Writes *hotel* after *my hotel*

Table 1 : example of a transcription of a multimodal session where the multimodal behavior of the user is “between” redundancy and complementarity.

Speech	If the recognized sentence contains the unique name of an object (i.e. "the Orsay museum"), set the salience of this object to 1.0. If the utterance contains only the value of a property of an object (i.e. "the museum"), increase the salience in the speech modality of all referencable object having the same property value (i.e. all the museums).
Gesture	Set the salience in the gesture modality as a function of the distance between the location of the object and the focus point of the recognized gesture.
Graphics	Set the salience in the graphics modality as a function of the distance between the location of the object and the center of the screen.
History	After the recognition of a command, the salience of objects referred to in this command is decreased by a forgetting factor.

Table 2: Informal definition of some of the rules used for updating the salience of objects as a function of multimodal behavior.