MultiModal semantic representation

Paul Mc Kevitt
School of Computing & Intelligent Systems
Faculty of Informatics
University of Ulster, Magee
Derry/Londonderry BT48 7JL, Northern Ireland.
E-mail: p.mckevitt@ulster.ac.uk

Abstract

Intelligent MultiMedia or MultiModal systems involve the computer processing, understanding and production of inputs and outputs from at least speech, text, and visual information in terms of semantic representations. One of the central questions for these systems is what form of semantic representation should be used, which of course goes back to the age old question of knowledge representation in artificial intelligence. When a system processes multimodal input it needs to map that input into the representation and vice-versa there needs to be a mapping out of the representation for multimodal output presentation. In addition, there are related issues of synchronisation of input/output and information fusion and coordination. Here, we look at current trends in multimodal semantic representation which are mainly XML- and frame- based, relate our experiences in the development of multimodal systems (CHAMELEON and CONFUCIUS) and conclude that *producer/consumer*, *intention (speech acts)*, *semantic-content*, and *timestamps* are four important components of any multimodal semantic representation. In addition, multimodal semantic representations depend on the task at hand, system architecture, will be necessary at different levels (media-independent and dependent) and will have numerous forms of representation. Semantic representations and content will need to provide for reference and spatial relations, two key recurring problems in multimodal systems.

1 Introduction

What distinguishes traditional MultiMedia from Intelligent MultiMedia or MultiModal Systems is that although both are concerned with text, voice, sound and video/graphics with possibly touch and virtual reality linked in, in the former the computer has little or no understanding of the meaning of what it is presenting. Intelligent MultiMedia or MultiModal systems involve the computer processing and understanding of perceptual signal and symbol input from at least speech, text and visual images, and then reacting to it, is much more complex and involves signal and symbol processing techniques from not just engineering and computer science but also artificial intelligence and cognitive science (Mc Kevitt 1994, 1995/96, Mc Kevitt et al. 2002). With IntelliMedia systems, people can interact in spoken dialogues with machines, querying about what is being presented and even their gestures and body language can be interpreted.

Although there has been much success in developing theories, models and systems in the areas of Natural Language Processing (NLP) and Vision Processing (VP) (Partridge 1991, Rich and Knight 1991) there has been little progress in integrating these two subareas of Artificial Intelligence (AI). In the beginning although the general aim of the field was to build integrated language and vision systems, few were, and these two subfields quickly arose. It is not clear why there has not already been much activity in integrating NLP and VP. Is it because of the long-time reductionist trend in science up until the recent emphasis on chaos theory, non-linear systems, and emergent behaviour? Or, is it because the people who have tended to work on NLP tend to be in other Departments, or of a different ilk, from those who have worked on VP? Dennett (1991, p. 57-58) says "Surely a major source of the widespread skepticism about "machine understanding" of natural language is that such systems almost never avail themselves of anything like a visual workspace in which to parse or analyze the input. If they did, the sense that they were actually understanding what they processed would be greatly heightened (whether or not it would still be, as some insist, an illusion). As it is, if a computer says, "I see what you mean" in response to input, there is a strong temptation to dismiss the assertion as an obvious fraud."

People are able to combine the processing of language and vision with apparent ease. In particular, people can use words to describe a picture, and can reproduce a picture from a language description. Moreover, people can exhibit this kind of behaviour over a very wide range of input pictures and language descriptions. Even more impressive is the fact that people can look at images and describe not just the image itself but a set of abstract emotions evoked by it. Although there are theories of how we process vision and language, there are few theories about how such processing is integrated. There have been large debates in Psychology and Philosophy with respect to the degree to which people store knowledge as propositions or pictures (Kosslyn and Pomerantz 1977, Pylyshyn 1973).

There are at least two advantages of linking the processing of natural languages to the processing of visual scenes. First, investigations into the nature of human cognition may benefit. Such investigations are being conducted in the fields of Psychology, Cognitive Science, and Philosophy. Computer implementations of integrated VP and NLP can shed light on how people do it. Second, there are advantages for real-world applications. The combination of two powerful technologies promises new applications: automatic production of speech/text from images; automatic production of images from speech/text; and the automatic interpretation of images with speech/text. The theoretical and practical advantages of linking natural language and vision processing have also been described in Wahlster (1988).

Early work for synthesizing simple text from images was conducted by Waltz (1975) who produced an algorithm capable of labelling edges and corners in images of polyhedra. The labelling scheme obeys a constraint minimisation criterion so that only sets of consistent labellings are used. The system can be expected to become 'confused' when presented with an image where two mutually exclusive but self-consistent labellings are possible. This is important because in this respect the program can be regarded as perceiving an illusion such as what humans see in the Necker cube. However, the system seemed to be incapable of any higher-order text descriptions. For example, it did not produce natural language statements such as "There is a cube in the picture."

A number of natural language systems for the description of image sequences have been developed (Herzog and Retz-Schmidt 1990, Neumann and Novak 1986). These systems can verbalize the behaviour of human agents in image sequences about football and describe the spatio-temporal properties of the behaviour observed. Retz-Schmidt (1991) and Retz-Schmidt and Tetzlaff (1991) describe an approach which yields plan hypotheses about intentional entities from spatio-temporal information about agents. The results can be verbalized in natural language. The system called REPLAI-II takes observations from image sequences as input. Moving objects from two-dimensional image sequences have been extracted by a vision system (Herzog et al. 1989) and spatio-temporal entities (spatial relations and events) have been recognised by an event-recognition system. A focussing process selects interesting agents to be concentrated on during a plan-recognition process. Plan recognition provides a basis for intention recognition and plan-failure analysis. Each recognised intentional entity is described in natural language. A system called SOCCER (André et al. 1988, Herzog et al. 1989) verbalizes real-world image sequences of soccer games in natural language and REPLAI-II extends the range of capabilities of SOCCER. Here, NLP is used more for annotation through text generation with less focus on analysis.

Maaβ et al. (1993) describe a system, called *Vitra Guide*, that generates multimodal route descriptions for computer assisted vehicle navigation. Information is presented in natural language, maps and perspective views. Three classes of spatial relations are described for natural language references: (1) topological relations (e.g. in, near), (2) directional relations (e.g. left, right) and (3) path relations (e.g. along, past). The output for all presentation modes relies on one common 3D model of the domain. Again, Vitra emphasizes annotation through generation of text, rather than analysis, and the vision module considers interrogation of a database of digitized road and city maps rather than vision analysis.

Some of the engineering work in NLP focusses on the exciting idea of incorporating NLP techniques with speech, touchscreen, video and mouse to provide advanced multimedia interfaces (Maybury 1993, Maybury and Wahlster 1998). Examples of such work are found in the ALFresco system which is a multimedia interface providing information on Italian Frescoes (Carenini et al. 1992 and Stock 1991), the WIP system that provides information on assembling, using, and maintaining physical devices like an expresso machine or a lawnmower (André and Rist 1993 and Wahlster et al. 1993) with more recent work on interactive presentations with an animated agent in PPP (Personalised Plan Presenter) (André et al. 1996, André and Rist 2000), AiA (Adaptive Communication Assistant for Effective Infobahn Access) (André and Rist 2001) and Miau (Multiple Internet Agents for User-Adaptive Decision Support) (André et al. 2000), and a multimedia interface which identifies objects and conveys route plans from a knowledge-based cartographic information system (Maybury 1991).

Others, developing general IntelliMedia platforms include *CHAMELEON* (Brøndsted et al. 1998, 2001) *SmartKom* (Reithinger 2001, Wahlster et al. 2001) *Situated Artificial Communicators* (Rickheit and Wachsmuth 1996), *Communicative Humanoids* (Thórisson 1996, 1997), *AESOPWORLD* (Okada 1996, 1997) and MultiModal Interfaces like *INTERACT* (Waibel et al. 1996). Other moves towards integration are reported in Denis and Carfantan (1993), Granström et al. (2002), Maybury (1997), Maybury and Wahlster (1998), Mc Kevitt (1994, 1995/96), Mc Kevitt et al. (2002) and Pentland (1993).

With the current proliferation of work in the area of Intelligent MultiMedia or MultiModal Systems one of the central questions people are asking is what is the correct semantic representation. And we must keep in mind of course that multimodal semantics not only applies to multimodal systems but also to efforts on semantic markup of the World Wide Web or *The Semantic Web* (see Berners-Lee et al. 2001).

2 MultiModal semantic representation

Detailed discussions on the nature and requirements of multimodal semantic representations are to be found in Romary (2001), Maybury (2001) and Bunt and Romary (2002). Chai et al. (2002) present their views on what such a semantics should contain. It is clear that a multimodal semantic representation must support interpretation and generation, any kind of multimodal input and output and a variety of semantic theories. The representation may contain architectural, environmental, and interactional information. Architectural information includes producer/consumer of the information, information confidence, and input/output devices. Environmental representation includes timestamps and spatial information. Interactional representation includes speaker/user's state.

Much of the work in MultiModal Systems chooses frames or XML to represent multimodal semantics. Frames are used in CHAMELEON, AESOPWORLD, REA (Cassell et al. 2000), Ymir (Thorisson 1996, 1997) and WordsEye (Coyne and Sproat 2001). The semantics can be localised as in CHAMELEON where the frames are stored in a central blackboard or distributed throughout various modules as in Ymir. XML-based representations are used in BEAT (Cassell et al. 2001), SmartKom (Wahlster et al. 2001) using M3L (MultiModal Markup Language), MIAMM using MMIL (MultiModal Interface Language) (Reithinger et al. 2002), MUST (Almeida et al. 2002) using MXML (MUST XML) and IMPROVISE (Zhou and Feiner 2001).

There are other multimodal systems using alternative specialised semantic representations. Ahn et al. (1996) and Bunt et al. (1998) use type theoretical logic within the DenK system, an electronic cooperative assistant, to represent domain knowledge, dialogue context, and a context-change theory of communication. Siskind (1995) uses event-logic truth conditions for simple spatial motion verbs in ABIGAIL which focusses on segmenting continuous motion pictures into distinct events and classifying those events into event types. Bailey et al. (1997) use x-schemas (eXecuting schemas) and f-structs (Feature-STRUCTures) representations which combine schemata representations with fuzzy set theory. They uses a formalism of Petri nets to represent x-schemas as a stable state of a system that consists of small elements which interact with each other when the system is moving from state to state. Narayanan et al. (1995) discuss the possibility of developing visual primitives for language primitives and use Schank's (1973) Conceptual Dependency (CD) theory in a 3D language visualisation system. As an alternative to symbolic representation methods for multimodal semantics there are also connectionist methods. Sales et al. (1996) in their *Neural State Machine* investigate Weightless Artificial Neural Network connectionist representations for grounding visual and linguistic representations. Feldman et al. (1996) in the L_0 project look at how a system can learn sentence-picture pairs. They started out using connectionist methods for grammar learning but then adopted a probabilistic framework which was thought to provide more versatile representations. Grumbach (1996) investigates how a hybrid connectionist model can be used to model implicit knowledge (e.g. sensori-motor associations) and explicit knowledge (e.g. a teacher giving verbal advice). Waibel et al. (1996) look at multimodal human computer interfaces with spoken dialogue, face recognition and gesture tracking with mainly neural network and statistical methods.

In addition to the various methods deployed for multimodal semantics within multimodal systems there are also moves from bodies, mainly industrial, to define markup languages for multimodal systems. SALT (Speech Application Language Tags) (2002) is an open standard attempt to augment existing XML-based markup languages in order to provide spoken access to many forms of content through a wide variety of devices, to promote multimodal interaction and to enable voice on the internet. The SALT specification language defines a set of lightweight tags as extensions to commonly used Web-based markup languages. VoiceXML (2002) arose from a need to define a markup language for over-the-telephone dialogues and at a time, 1999, when many pieces of the Web infrastructure as we know it today had not matured. There are also additional semantic markup languages within the XML family of the WorldWideWeb Consortium (W3C) such as Ontology Web Language (OWL) published by the W3C's Web Ontology Working Group (OWL 2002). OWL is a derivative of DAML+OIL (DARPA Agent Markup Language, Ontology Interchange Language) Web Ontology Language (DAML+OIL 2002) and builds upon the Resource Description Framework (RDF). Also, relevant is the fact that W3C has a Working Group on Multimodal Interaction looking at Multimodal interaction on the web with specific focus on a markup specification for synchronisation across various modalities and devices with a wide range of capabilities (W3C-MMI 2002).

3 MultiModal experiences: CHAMELEON and CONFUCIUS

We have had experience with developing two MultiModal systems, CHAMELEON and CONFUCIUS and each system has its own requirements in terms of MultiModal semantic representation.

3.1 CHAMELEON

CHAMELEON has a distributed architecture of communicating agent modules processing inputs and outputs from different modalities and each of which can be tailored to a number of application domains. The process synchronisation and intercommunication for CHAMELEON modules is performed using the DACS (Distributed Applications Communication System) Inter Process Communication (IPC) software (see Fink et al. 1996) which enables CHAMELEON modules to be glued together and distributed across a number of servers. Presently, there are ten software modules in CHAMELEON: blackboard, dialogue manager, domain model, gesture recogniser, laser system, microphone array, speech recogniser, speech synthesiser, natural language processor (NLP), and Topsy as shown in Figure 1. More detail on CHAMELEON can be found in Brøndsted et al. 1998, 2001).

An initial application of CHAMELEON is the *IntelliMedia WorkBench* which is a hardware and software platform as shown in Figure 2. One or more cameras and lasers can be mounted

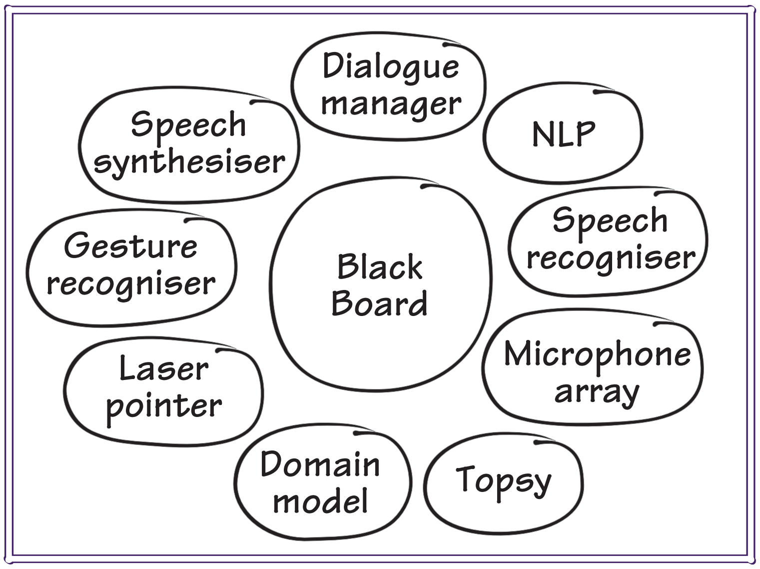

Figure 1: Architecture of CHAMELEON

in the ceiling, microphone array placed on the wall and there is a table where things (objects, gadgets, people, pictures, 2D/3D models, building plans, or whatever) can be placed. The current domain is a *Campus Information System* which at present gives information on the architectural and functional layout of a building. 2-dimensional (2D) architectural plans of the building drawn on white paper are laid on the table and the user can ask questions about them. Presently, there is one static camera which calibrates the plans on the table and the laser, and interprets the user's pointing while the system points to locations and draws routes with a laser. Inputs are simultaneous speech and/or pointing gestures and outputs are synchronised speech synthesis and pointing. We currently run all of CHAMELEON on a standard Intel pentium computer which handles input for the Campus Information System in real-time.

3.2 Frame semantics

CHAMELEON's *blackboard* stores semantic representations produced by each of the other modules and keeps a history of these over the course of an interaction. All modules communicate through the exchange of semantic representations with each other or the blackboard. The meaning of interactions over the course of a MultiModal dialogue is represented using a frame semantics with frames in the spirit of Minsky (1975). The intention is that all modules in the system can produce and read frames. Frames are coded in CHAMELEON with messages built as predicate-argument structures following a BNF definition. The frame semantics was first presented in Mc Kevitt and Dalsgaard (1997). Frames represent some crucial elements such as *module*, *input/output*, *intention*, *location*, and *timestamp*. Module is simply the name of the module producing the frame (e.g. NLP). Inputs are the input recognised whether spoken (e.g. "Show me Hanne's office") or gestures (e.g. pointing coordinates) and outputs the intended output whether spoken (e.g. "This is Hanne's office.") or gestures (e.g. pointing coordinates). Timestamps can include the times a given module commenced and terminated processing and the time a frame was written on the blackboard. The frame semantics also includes representations for two key phenomena in language/vision integration: reference and spatial relations.

Frames can be grouped into three categories: (1) *input*, (2) *output* and (3) *integration*. Input frames are those which come from modules processing perceptual input, output frames are those produced by modules generating system output and integration frames are integrated meaning representations constructed over the course of a dialogue (i.e. all other frames). Here, we shall discuss frames with a focus more on frame semantics than on frame syntax and in fact the actual

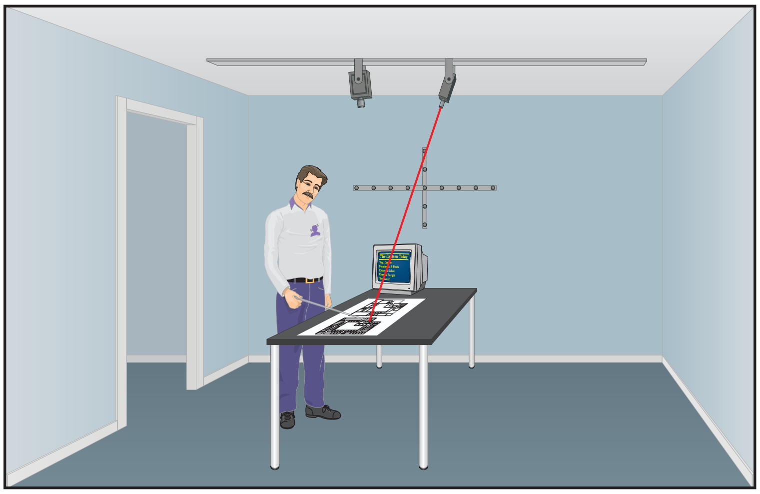

Figure 2: Physical layout of the IntelliMedia WorkBench

coding of frames as messages within CHAMELEON has a different syntax.

3.2.1 Input frames

An input frame takes the general form:

[MODULE
INPUT: input
INTENTION: intention-type
TIME: timestamp]

where MODULE is the name of the input module producing the frame, INPUT can be at least UTTERANCE or GESTURE, *input* is the utterance or gesture and *intention-type* includes different types of utterances and gestures. An utterance input frame can at least have intention-type (1) query?, (2) instruction! and (3) declarative. An example of an utterance input frame is:

[SPEECH-RECOGNISER
UTTERANCE: (Point to Hanne's office)
INTENTION: instruction!
TIME: timestamp]

A gesture input frame is where *intention-type* can be at least (1) pointing, (2) mark-area, and (3) indicate-direction. An example of a gesture input frame is:

[GESTURE
GESTURE: coordinates (3, 2)
INTENTION: pointing
TIME: timestamp]

3.2.2 Output frames

An output frame takes the general form:

[MODULE
INTENTION: intention-type
OUTPUT: output
TIME: timestamp]

where MODULE is the name of the output module producing the frame, *intention-type* includes different types of utterances and gestures and OUTPUT is at least UTTERANCE or GESTURE. An utterance output frame can at least have intention-type (1) query? (2) instruction!, and (3) declarative. An example utterance output frame is:

[SPEECH-SYNTHESIZER
INTENTION: declarative
UTTERANCE: (This is Hanne's office)
TIME: timestamp]

A gesture output frame can at least have intention-type (1) description (pointing), (2) description (route), (3) description (mark-area), and (4) description (indicate-direction). An example gesture output frame is:

[LASER
INTENTION: description (pointing)
LOCATION: coordinates (5, 2)
TIME: timestamp]

3.2.3 Integration frames

Integration frames are all those other than input/output frames. An example utterance integration frame is:

[NLP
INTENTION: description (pointing)
LOCATION: office (tenant Hanne) (coordinates (5, 2))
UTTERANCE: (This is Hanne's office)
TIME: timestamp]

Things become even more complex with the occurrence of references and spatial relationships:

[MODULE
INTENTION: intention-type
LOCATION: location
LOCATION: location
LOCATION: location
SPACE-RELATION: beside
REFERENT: person
LOCATION: location
TIME: timestamp]

An example of such an integration frame is:

[DOMAIN-MODEL
INTENTION: query? (who)

LOCATION: office (tenant Hanne) (coordinates (5, 2))
LOCATION: office (tenant Jørgen) (coordinates (4, 2))
LOCATION: office (tenant Børge) (coordinates (3, 1))
SPACE-RELATION: beside
REFERENT: (person Paul-Dalsgaard)
LOCATION: office (tenant Paul-Dalsgaard) (coordinates (4, 1))
TIME: timestamp]

We have reported complete blackboard histories for the instruction "Point to Hanne's office" and the query "Whose office is this?" + [pointing] (exophoric/deictic reference) in Mc Kevitt and Dalsgaard (1997) and Brøndsted et al. (1998). With respect of spatial relations we derive all the frames appearing on the blackboard for the example: "Who's in the office beside him?' in Mc Kevitt (2000).

To summarise, in CHAMELEON and the IntelliMedia Workbench we have found that *producer/consumer*, *intention (speech acts)*, *semantic-content*, and *timestamps* are four important components of any multimodal semantic representation. With respect of multimodal semantic-content there is a requirement of representing two key elements of multimodal systems: reference and spatial relations.

4 Seanchaí

Within an intelligent multimedia storytelling platform called Seanchaí we are interested in generating 3D animation automatically. Seanchaí will perform multimodal storytelling generation, interpretation and presentation and consists of *Homer*, a storytelling generation module, and CONFUCIUS, a storytelling interpretation and presentation module (see Figure 3). The output of the former module could be fed as input to the latter. Homer focuses on natural language story generation. It will receive two types of input from the user, (1) either the beginning or the ending of a story in the form of a sentence, and (2) stylistic specifications, and outputs natural language stories; and CONFUCIUS focuses on story interpretation and multimodal presentation. It receives input natural language stories or (play/movie) scripts and presents them with 3D animation, speech and non-speech audio.

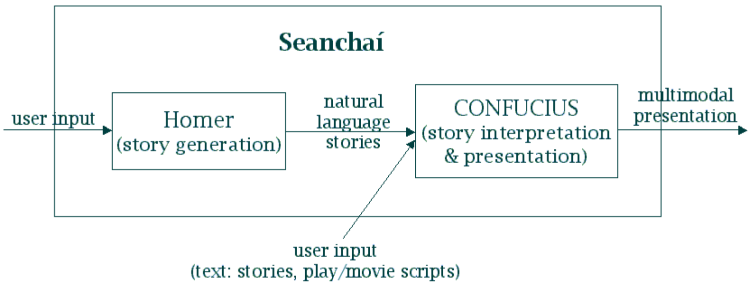

Figure 3: Intelligent multimodal storytelling platform – *Seanchaí*

The knowledge base and its visual knowledge semantic representation are used in CONFUCIUS (see Figure 4), and they could also be adopted in other vision and natural language processing integration applications. The dashed part in the figure includes the prefabricated objects such as characters, props, and animations for basic activities, which will be used in the *Animation generation* module. When the input is a story, it will be transferred to a script by the *script writer*, then parsed by the *script parser* and the *natural language processing* module respectively. The modules for *Natural Language Processing (NLP)*, *Text to Speech (TTS)* and *sound effects* operate in parallel. Their outputs will be fused at *code combination*, which generates a holistic 3D world representation including animation, speech and sound effects. NLP

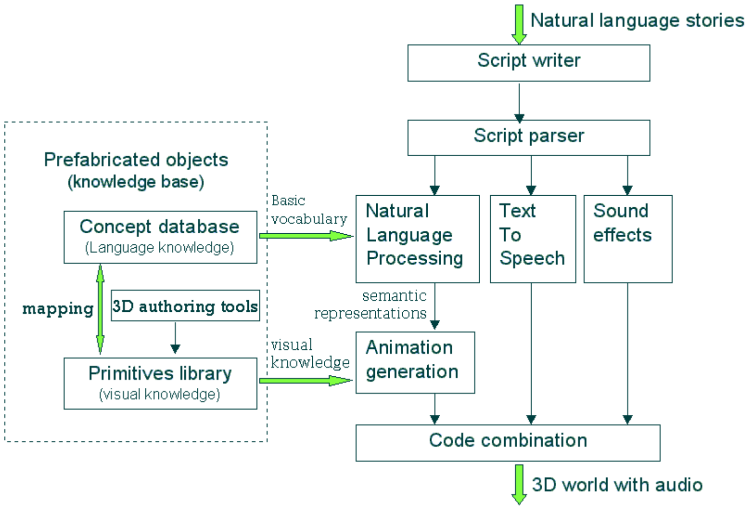

Figure 4: System architecture of CONFUCIUS

will be performed using Gate and WordNet, TTS will be performed using Festival or Microsoft Whistler, VRML (Virtual Reality Modelling Language) will be used to model the story 3D virtual world, and visual semantics is represented using a Prolog-like formalism.

4.1 Visual knowledge representation

Existing multimodal semantic representations within various intelligent multimedia systems may represent the general organisation of semantic structure for various types of inputs and outputs and are usable at various stages such as media fusion and pragmatic aspects. However, there is a gap between high-level general multimodal semantic representation and lower-level representation that is capable of connecting meanings across modalities. Such a lower-level meaning representation, which links language modalities to visual modalities, is proposed in Ma and Mc Kevitt (2003). Figure 5 illustrates the multimodal semantic representation of CONFUCIUS. It is composed of language, visual and non-speech audio modalities. Between the multimodal semantics and each specific modality there are two levels of representation: one is a high-level multimodal semantic representation which is media-independent, the other is an intermediate level media-dependent representation. CONFUCIUS will use an XML-based representation for high-level multimodal semantics and an extended predicate-argument representation for intermediate representation which connects language with visual modalities as shown in Figure 5. Our visual semantics decomposition method is at the intermediate representation level (see Ma and Mc Kevitt 2003). It is suitable for implementation in the 3D graphic modelling language VRML. It will be translated to VRML code by a Java program in CONFUCIUS. We also plan to include non-speech audio in the media-dependent and media-independent semantic representations.

The predicate-argument format we apply to represent verb semantics has a Prolog-inspired nomenclature. Each non-atomic action is defined by one or more subgoals, and the name of every goal/subgoal reveals its purpose and effect. Primitives 1 through 14 are basic primitive actions in our framework (Figure 6). We do not claim that these fourteen cover all the necessary primitives needed in modelling observable verbs. 13[1] and 14[2] are actually not primitive actions, but they are necessary in processing complex space displacement. In the first twelve primitives, 1-3 describe position movement, 4 and 5 concern orientation changes, 6-9 focus on alignment, 10

[1]As is the convention in the programming language Prolog, arguments can be replaced by an underscore if they are undetermined.

[2]ungroup element x from a list which contains it. yList is the rest of the list after deleting x from the original list. This is also a basic list operation in Prolog.

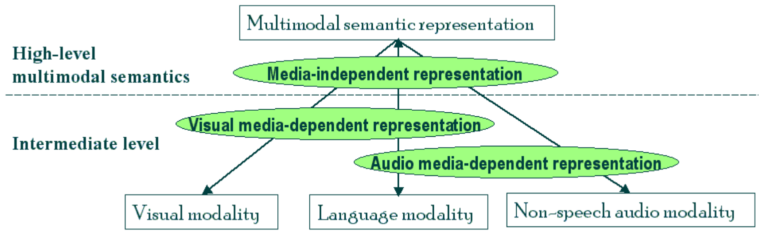

Figure 5: MultiModal semantic representation in CONFUCIUS

is a composite action (not atomic) composed by lower level primitives, and 11, 12 concern size (shape) changes. Figure 7 illustrates the hierarchical structure of the twelve primitives. Higher level actions are defined by lower level ones. For instance, alignment operations are composed by `move()` and/or `moveTo()` predicates. Definitions of the primitives are given in Ma and Mc Kevitt (2003).

```
 1) move(obj, xInc, yInc, zInc)
 2) moveTo(obj, loc)
 3) moveToward(obj, loc, displacement)
 4) rotate(obj, xAngle, yAngle, zAngle)
 5) faceTo(obj1, obj2)
 6) alignMiddle(obj1, obj2, axis)
 7) alignMax(obj1, obj2, axis)
 8) alignMin(obj1, obj2, axis)
 9) alignTouch(obj1, obj2, axis)
10) touch(obj1, obj2, axis)  ; for the relation of support and contact
11) scale(obj, rate)  ; scale up/down, change size
12) squash(obj, rate, axis)  ; squash or lengthen an object
13) group(x, [y|_], newObj)
14) ungroup(xyList, x, yList)
```

Figure 6: Basic predicate-argument primitives within CONFUCIUS

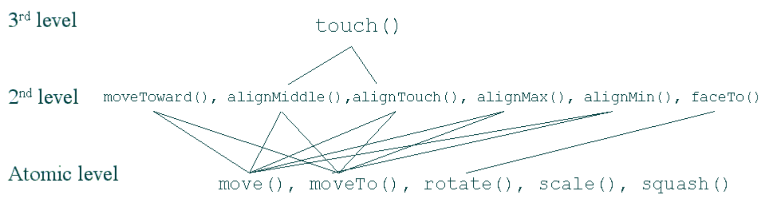

Figure 7: Hierarchical structure of CONFUCIUS' primitives

The predicate-argument primitives can be used to provide definitions of visual semantics of verbs. For example,

Example 1, *jump*[3] :

```
jump(x):-
```

[3]Semantic constraint – declare an instance of the type 'Animal'. Metaphor usage of vegetal or inanimate characters is not considered here.

```
        type(x, Animal),
        move(x.feet, _, HEIGHT, _),
        move(x.body, _, HEIGHT, _),
        move(x.feet, _, -HEIGHT, _).
```

Example 2, *call*:

– as in "A is calling B" (verb tense is not considered here because it is at sentence level rather than word level). This is one word-sense of call where calling is conducted by telephone. Here is the definition of one word-sense of call which is at the first level of the visual semantic verb representation hierarchy:

```
call(a):-
        type(a, Person),
        type(tel, Telephone),
        pickup(a, tel.receiver, a.leftEar),
        dial(a, tel.keypad),
        speak(a, tel.receiver),
        putdown(a, tel.receiver, tel.set).
```

Further examples are given in Ma and Mc Kevitt (2003).

To summarise, in CONFUCIUS we have found that as in CHAMELEON higher-level media-independent semantic representations will be important in forms such as XML and frames but also that intermediate-level media-dependent representations will be necessary in order to represent fully correspondences between modalities.

5 Discussion

Our experience with MultiModal semantic representation is that the representations required are dependent on the applications at hand and also MultiModal system architectures. This is also clear from the discussions found in Romary (2001), Maybury (2001) and Bunt and Romary (2002). There are requirements for higher-level media-independent representations but also lower-level more media-dependent representations. We argue that producer/consumer, intention (speech acts), semantic-content, and timestamps are four important components of any higher-level multimodal semantic representation.

Many of the requirements in multimodal semantic representation come from the need to integrate information from different modalities. In terms of language and vision integration there are requirements for mapping the language and visual information into semantic components which can be fused and integrated and will be necessary for answering queries such as "Whose office is this?" In terms of language and computer graphics integration there are requirements for determining the visual meaning of language actions (verbs) so that for example, language can be mapped into graphical presentations automatically. So for example with the verb "close" there could be three visual definitions: closing of a normal door (rotation on y axis), closing of a sliding door (moving on x axis), or closing of a rolling shutter door (a combination of rotation on x axis and moving on y axis).

Two key problems in language and vision integration are reference (see Brøndsted 1999, Kievet et al. 2001) and spatial relations (see Mc Kevitt 2000, Zelinsky-Wibbelt 1993), i.e. in multimodal systems there are regular deictic references to the visual context and also numerous spatial relations. Hence, it is a necessary requirement for adequate semantic-content representations to incorporate mechanisms for representing spatial relations and reference.

6 Conclusion and future work

Although traditional and Intelligent MultiMedia or MultiModal Systems are both concerned with text, voice, sound and video/graphics, with the former the computer has little or no understanding of the meaning of what it is presenting and this is what distinguishes the two.

With the current proliferation of multimodal systems the question that everyone is asking is what is the correct multimodal meaning representation. From our experience in developing two multimodal systems, one which integrates the processing of spoken dialogue and vision for both input and output (CHAMELEON) and one which translates text stories into multimodal presentations with 3D graphics, spoken dialogue and non-speech audio (CONFUCIUS) we conclude that multimodal semantic representation: (1) depends on the task at hand, (2) depends on the system architecture, (3) will be necessary at different levels (media-independent and dependent) (4) will have at least the following four important components: *producer/consumer*, *intention (speech acts)*, *semantic-content*, and *timestamps* (5) will have many forms of representation such as frames, XML, formal logics, event-logic truth conditions, X-schemas and f-structs or connectionist models. With respect of multimodal semantic-content there is a requirement of representing two key elements of multimodal systems: reference and spatial relations. With respect of multimodal system architectures there are interesting questions as to where multimodal semantic representations lie in systems and whether all the semantics is contained in one single blackboard (CHAMELEON) or distributed throughout the system (Ymir and SmartKom).

Future work will involve experimenting with various semantic representations and architectures with numerous applications and as we have found with knowledge representation in artificial intelligence it may be the case that no single representation is the correct one but more significant will be how we use the representation and what can be achieved with it in terms of multimodality.

References

Ahn, R., R.J. Beun, T. Borghuis, H.C. Bunt and C. van Overveld (1995) The DENK architecture: a fundamental approach to user-interfaces. In *Integration of natural language and vision processing (Vol. I): computational models and systems*, P. Mc Kevitt (Ed.), 267-281. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Almeida, L., I. Amdal, N. Beires, M. Boualem, L. Boves, E. den Os, P. Filoche, R. Gomes, J.E. Knudsen, K. Kvale, J. Rugelbak, C. Tallec and N. Warakagoda (2002) Implementing and evaluating a multimodal and multilingual tourist guide. In *Proc. of the International CLASS Workshop on Natural, Intelligent and Effective interaction in MultiModal Dialogue Systems*, Copenhagen, Denmark, 28-29 June.

André, Elisabeth, G. Herzog, and T. Rist (1988) On the simultaneous interpretation of real-world image sequences and their natural language description: the system SOCCER. In *Proceedings of the 8th European Conference on Artificial Intelligence*, 449-454, Munich, Germany.

André, Elisabeth and Thomas Rist (1993) The design of illustrated documents as a planning task. In *Intelligent multimedia interfaces*, M. Maybury (Ed.), 75-93. Menlo Park, CA: AAAI Press.

André, E., J. Müller and T. Rist (1996) The PPP persona: a multipurpose animated presentation agent. In *Advanced Visual Interfaces*, T. Catarci, M.F. Constabile, S. Levialdi and G. Santucci (Eds.), 245-247. New York, USA: ACM Press.

André, E. and T. Rist (2000) Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems. In *Proc. of the Second International Conference on Intelligent User Interfaces (IUI 2000)*, Los Angeles, CA, USA, 1-8.

André, E., T. Rist, S. van Mulken, M. Klesen and S. Baldes (2000) The automated design of believeable dialogues for animated presentation teams. In *Embodied conversational agents*, J. Cassell, J. Sullivan, S. Prevost and E. Churchill (Eds.), 220-255. Cambridge, MA: The MIT Press.

André, E. and T. Rist (2001) Controlling the behaviour of animated presentation agents in the interface: scripting vs. instructing. In *AI Magazine*, 22(4), 53-66.

Bailey, D., J. Feldman, S. Narayanan & G. Lakoff (1997) Modeling embodied lexical development. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (CogSci97), 19-24*, Stanford, CA, USA.

Berners-Lee, T., J. Hendler and O. Lassila (2001) The semantic web. In *Scientific American*, May.

Brøndsted, T. (1999) Reference problems in CHAMELEON. In *Proc. of the ESCA Tutorial and Research Workshop on Interactive Dialogue in MultiModal Systems (IDS-99)*, Paul Dalsgaard, Paul Heisterkamp and Chin-Hui Lee (Eds.), 133-136. Kloster Irsee, Germany, June.

Brøndsted, T., P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund & K.G. Olesen (1998) *A platform for developing Intelligent MultiMedia applications.* Technical Report R-98-1004, Center for PersonKommunikation (CPK), Institute of Electronic Systems (IES), Aalborg University, Denmark, May.

Brøndsted, T., P. Dalsgaard, L.B. Larsen, M. Manthey, P. Mc Kevitt, T.B. Moeslund & K.G. Olesen (2001) The IntelliMedia WorkBench - an Environment for Building Multimodal Systems. In *Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98, Tilburg, The Netherlands, January 1998, Selected Papers*, H. Bunt and R.J. Beun (Eds.), 217-233. Lecture Notes in Artificial Intelligence (LNAI) series, LNAI 2155, Berlin, Germany: Springer Verlag.

Bunt, H.C., R. Ahn, R.J. Beun, T. Borghuis and C. van Overveld (1998) Multimodal cooperation with the DENK system. In *Multimodal Human-Computer Communication*, H.C. Bunt, R.J. Beun and T. Borghuis (Eds.), 1-12. Berlin, Germany: Springer-Verlag.

Bunt, H. and L. Romary (2002) Towards multimodal content representation. In *International standards of terminology and language resources management*, LREC 2002, Las Palmas, Spain.

Carenini, G., F. Pianesi, M. Ponzi and O. Stock (1992) *Natural language generation and hypertext access.* IRST Technical Report 9201-06, Instituto Per La Scientifica E Tecnologica, Loc. Pant e Di Povo, I-138100 Trento, Italy.

Cassell, J., J. Sullivan, S. Prevost and E. Churchill (Eds.) (2000) *Embedded conversational agents.* Cambridge, MA: MIT Press.

Cassell, J., H. Vilhjalmsson and T. Bickmore (2001) BEAT: the behaviour expression animation toolkit. In *SIGGRAPH 2001 Conference Proceedings*, Los Angeles, August 12-17, 477-486.

Chai, J. S. Pan and M.X. Zhou (2002) MIND: semantics based multimodal interpretation framework. In *Proc. of the International CLASS Workshop on Natural, Intelligent and Effective interaction in MultiModal Dialogue Systems*, Copenhagen, Denmark, 28-29 June.

Coyne, B & R. Sproat (2001) WordsEye: an automatic text-to-scene conversion system. In *AT&T Labs. Computer Graphics Annual Conference, SIGGRAPH 2001 Conference Proceedings*, Los Angeles, Aug 12-17, 487-496.

DAML+OIL (2002) *DAML+OIL reference description.* http://www.w3.org/TR/daml+oil-reference.

Denis, M. and M. Carfantan (Eds.) (1993) *Images et langages: multimodalité et modelisation cognitive.* Actes du Colloque Interdisciplinaire du Comité National de la Recherche Scientifique, Salle des Conférences, Siège du CNRS, Paris, April.

Dennett, Daniel (1991) *Consciousness explained.* Harmondsworth: Penguin.

Feldman, J., G. Lakoff, D. Bailey, S. Narayanan, T. Regier and A. Stolcke (1996) L_0 – the first five years of an automated language acquisition project. In *Integration of natural language and vision processing (Vol. III): theory and grounding representations*, P. Mc Kevitt (Ed.), 205-231. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fink, G.A., N. Jungclaus, H. Ritter, and G. Sagerer (1995) A communication framework for heterogeneous distributed pattern analysis. In *Proc. International Conference on Algorithms and Applications for Parallel Processing*, V. L. Narasimhan (Ed.), 881-890. IEEE, Brisbane, Australia.

Granström, Björn, David House and Inger Karlsson (Eds.) (2002) *Multimodality in language and Speech systems.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Grumbach, A. (1996) Grounding symbols into perceptions. In *Integration of natural language and vision processing (Vol. III): theory and grounding representations*, P. Mc Kevitt (Ed.), 233-248. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Herzog, G. and G. Retz-Schmidt (1990) Das System SOCCER: Simultane Interpretation und naturalichsprachliche Beschreibung zeitveranderlicher Szenen. In *Sport und Informatik*, J. Perl (Ed.), 95-119. Schorndorf: Hofmann.

Herzog, G., C.-K. Sung, E. Andre, W. Enkelmann, H.-H. Nagel, T. Rist, and W. Wahlster (1989) Incremental natural language description of dynamic imagery. In *Wissenbasierte Systeme. 3. Internationaler GI-Kongress*, C. Freksa and W. Brauer (Eds.), 153-162. Berlin: Springer-Verlag.

Kievet, L, P. Piewek, R. Jan-Beun and H. Bunt (2001) Multimodal cooperative resolution of referential expressions in the DENK system. In *Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98, Tilburg, The Netherlands, January 1998, Selected Papers*, H.C. Bunt and R.J. Beun (Eds.), 197-214. Lecture Notes in Artificial Intelligence (LNAI) series, LNAI 2155, Berlin, Germany: Springer-Verlag.

Kosslyn, S.M. and J.R. Pomerantz (1977) Imagery, propositions and the form of internal representations. In *Cognitive Psychology*, 9, 52-76.

Ma, Minhua and Paul Mc Kevitt (2003) Semantic representation of events in 3D animation. In *Proc. of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Harry Bunt, Reinhard Muskens and Elias Thiesse (Eds.). Tilburg University, Tilburg, The Netherlands, January.

Maaβ, Wolfgang, Peter Wizinski, Gerd Herzog (1993) *VITRA GUIDE: Multimodal route descriptions for computer assisted vehicle navigation.* Bereich Nr. 93, Universitat des Saarlandes, FB 14 Informatik IV, Im Stadtwald 15, D-6600, Saarbrucken 11, Germany, February.

Maybury, Mark (1991) Planning multimedia explanations using communicative acts. In *Proceedings of the Ninth American National Conference on Artificial Intelligence (AAAI-91)*, Anaheim, CA, July 14-19.

Maybury, Mark (Ed.) (1993) *Intelligent multimedia interfaces.* Menlo Park, CA: AAAI Press.

Maybury, Mark (Ed.) (1997) *Intelligent multimedia information retrieval.* Menlo Park, CA: AAAI/MIT Press.

Maybury, M. (2001) *Would you build your dream house without a blueprint?, Working group on software architectures for MultiModal Systems (WG 3).* International Seminar on Coordination and Fusion in MultiModal Interaction, Schloss Dagstuhl International Conference and Research Center for Computer Science, Wadern, Saarland, Germany, 29 October - 2 November. (http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/WG_3_The_Architecture_Dream_Team/index.html).

Maybury, Mark and Wolfgang Wahlster (Eds.) (1998) *Readings in intelligent user interfaces.* Los Altos, CA: Morgan Kaufmann Publishers.

Mc Kevitt, Paul (1994) Visions for language. In *Proceedings of the Workshop on Integration of Natural Language and Vision processing*, Twelfth American National Conference on Artificial Intelligence (AAAI-94), Seattle, Washington, USA, August, 47-57.

Mc Kevitt, Paul (Ed.) (1995/1996) *Integration of Natural Language and Vision Processing (Vols. I-IV).* Dordrecht, The Netherlands: Kluwer-Academic Publishers.

Mc Kevitt, Paul (2000) CHAMELEON meets spatial cognition. In *Spatial cognition*, Sean O Nuallain (Ed.), 149-170. US: John Benjamins, Also in, Proceedings of MIND-III: The Annual Conference of the Cognitive Science Society of Ireland, Theme: Spatial Cognition, Mary Hegarty and Seán Ó Nualláin (Eds.), Part II, 70-87. Dublin City University (DCU), Dublin, Ireland, August, 1998.

Mc Kevitt, Paul and Paul Dalsgaard (1997) A frame semantics for an IntelliMedia TourGuide. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), Volume 1*, 104-111. University of Ulster, Magee, Derry, Northern Ireland, September.

Mc Kevitt, Paul, Seán Ó Nualláin and Conn Mulvihill (Eds.) (2002) *Language, vision and music.* Amsterdam, The Netherlands: John Benjamins Publishing Co..

Minsky, M. (1975) A Framework for representing knowledge. In *Readings in knowledge representation*, R. Brachman and H. Levesque (Eds.), 245-262. Los Altos, CA: Morgan Kaufmann.

Narayanan, S., D. Manuel, L. Ford, D. Tallis & M. Yazdani (1995) Language visualisation:

applications and theoretical foundations of a primitive-based approach. In *Integration of Natural Language and Vision Processing (Volume II): Intelligent Multimedia*, P. Mc Kevitt (Ed.), 143-163. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Neumann, B. and H.-J. Novak (1986) NAOS: Ein System zur naturalichsprachlichen Beschreibung zeitveranderlicher Szenen. In *Informatik. Forschung und Entwicklung*, 1(1): 83-92.

Okada, Naoyuki (1996) Integrating vision, motion and language through mind. In *Integration of Natural Language and Vision Processing, Volume IV, Recent Advances*, Mc Kevitt, Paul (ed.), 55-80. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Okada, Naoyuki (1997) Integrating vision, motion and language through mind. In *Proceedings of the Eighth Ireland Conference on Artificial Intelligence (AI-97), Volume 1*, 7-16. University of Ulster, Magee, Derry, Northern Ireland, September.

OWL (2002) *Feature synopsis for OWL Lite and OWL*. http://www.w3.org/TR/WD-owl-features-2020729/.

Partridge, Derek (1991) *A new guide to Artificial Intelligence*. Norwood, New Jersey: Ablex Publishing Corporation.

Pentland, Alex (Ed.) (1993) *Looking at people: recognition and interpretation of human action*. IJCAI-93 Workshop (W28) at The 13th International Conference on Artificial Intelligence (IJCAI-93), Chambéry, France, EU, August.

Pylyshyn, Zenon (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. In *Psychological Bulletin*, 80, 1-24.

Reithinger, N. (2001) *Media coordination in SmartKom*. International Seminar on Coordination and Fusion in MultiModal Interaction, Schloss Dagstuhl International Conference and Research Center for Computer Science, Wadern, Saarland, Germany, 29 October - 2 November. (http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/Media_Coordination_In_SmartKom/index.html).

Reithinger, N., C. Lauer and L. Romary (2002) MIAMM - Multidimensional information access using multiple modalities. In *Proc. of the International CLASS Workshop on Natural, Intelligent and Effective interaction in MultiModal Dialogue Systems*, Copenhagen, Denmark, 28-29 June.

Retz-Schmidt, Gudala (1991) Recognizing intentions, interactions, and causes of plan failures. In *User Modelling and User-Adapted Interaction*, 1: 173-202.

Retz-Schmidt, Gudala and Markus Tetzlaff (1991) *Methods for the intentional description of image sequences*. Bereich Nr. 80, Universitat des Saarlandes, FB 14 Informatik IV, Im Stadtwald 15, D-6600, Saarbrucken 11, Germany, EU, August.

Rich, Elaine and Kevin Knight (1991) *Artificial Intelligence*. New York: McGraw-Hill.

Rickheit, Gert and Ipke Wachsmuth (1996) Collaborative Research Centre "Situated Artificial Communicators" at the University of Bielefeld, Germany. In *Integration of Natural Language and Vision Processing, Volume IV, Recent Advances*, Mc Kevitt, Paul (ed.), 11-16. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Romary, L. (2001) *Working group on multimodal meaning representation (WG 4)*. International Seminar on Coordination and Fusion in MultiModal Interaction, Schloss Dagstuhl International Conference and Research Center for Computer Science, Wadern, Saarland, Germany, 29 October - 2 November. (http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/WG_4-_Multimodal_Meaning_Representation/index.html).

Sales, N.J., R.G. Evans and I. Aleksander (1996) Successful naive representation grounding. In *Integration of natural language and vision processing (Vol. III): computational models and systems*, P. Mc Kevitt (Ed.), 185-204. Dordrecht, The Netherlands: Kluwer Academic Publishers.

SALT (2002) *http://www.saltforum.org.* .

Schank, R.C. (1973) *The fourteen primitive actions and their inferences*. Memo AIM-183, Stanford Artificial Intelligence Laboratory, Stanford, CA, USA.

Siskind, J.M. (1995) Grounding language in perception. In *Integration of Natural Language and Vision Processing (Volume I): Computational Models and Systems*, P. Mc Kevitt (Ed.), 207-227. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stock, Oliviero (1991) Natural language and exploration of an information space: the ALFresco Interactive system. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, 972-978, Darling Harbour, Sydney, Australia, August.

Thórisson, Kris R. (1996) *Communicative humanoids: a computational model of psychosocial dialogue skills*. Ph.D. thesis, Massachusetts Institute of Technology.

Thórisson, Kris R. (1997) Layered action control in communicative humanoids. In *Proceedings of Computer Graphics Europe '97*, June 5-7, Geneva, Switzerland.

VoiceXML (2002) *http://www.voicexml.org.* .

Wahlster, Wolfgang (1988) *One word says more than a thousand pictures: On the automatic verbalization of the results of image sequence analysis.* Bereich Nr. 25, Universitat des Saarlandes, FB 14 Informatik IV, Im Stadtwald 15, D-6600, Saarbrucken 11, Germany, February.

Wahlster, Wolfgang, Elisabeth André, Wolfgang Finkler, Hans-Jurgen Profitlich, and Thomas Rist (1993) Plan-based integration of natural language and graphics generation. In *Artificial Intelligence, Special issue on natural language generation*, 63, 387-427.

Wahlster, W., N. Reithinger and A. Blocher (2001) SmartKom: towards multimodal dialogues with anthropomorphic interface agents. In *Proceedings of The International Status Conference: Lead Projects, "Human-Computer Interaction"*, G. Wolf and G. Klein (Eds.), 23-34. Berlin, Germany: Deutsches Zentrum für Luft- und Raumfahrt.

Waltz, David (1975) Understanding line drawings of scenes with shadows. In *The psychology of computer vision*, Winston, P.H. (Ed.), 19-91. New York: McGraw-Hill.

W3C-MMI (2002) *http://www.w3.org/2002/mmi/.* .

Zelinsky-Wibbelt, Cornelia (Ed.) (1993) *The semantics of prepositions: from mental Processing to natural language processing (NLP 3).* Berlin, Germany: Mouton de Gruyter.

Zhou, M.X. and S. Feiner (2001) IMPROVISE: automated generation of animated graphics for coordinated multimeida presentations. In *Advances in Cooperative Multimodal Communication: Second International Conference, CMC'98, Tilburg, The Netherlands, January 1998, Selected Papers*, H.C. Bunt and R.J. Beun (Eds.), 43-63. Lecture Notes in Artificial Intelligence (LNAI) series, LNAI 2155, Berlin, Germany: Springer-Verlag.