In: Stock, O., Zancanaro, M. (eds.): Multimodal Intelligent Information Presentation. Series Text, Speech and Language Technology, Vol 27. Kluwer Academic. pp. 325-340

FUSION AND COORDINATION FOR MULTIMODAL INTERACTIVE INFORMATION PRESENTATION

Roadmap, Architecture, Tools, Semantics

Harry Bunt[1], Michael Kipp[2], Mark T. Maybury[3], and Wolfgang Wahlster[2]

ABSTRACT

Users require more effective and efficient means of interaction with increasingly complex information and new interactive devices. This document summarizes the results of the international Dagstuhl Seminar on Coordination and Fusion in Multimodal Interaction that took place at Schloss Dagstuhl in Germany October 27 through November 2, 2001[1]. We first outline a research roadmap in the near and long term. Next we describe requirements and an abstract architecture for this class of systems. We then detail requirements for semantic representations and languages necessary to enable these systems. Finally, we describe data, annotation methodologies and tools necessary to further advance the field. We conclude with a recommended action plan for forward progress in the community.

1. 0 ROADMAP

Figure 1 illustrates the roadmap in the near term, from 2002-2005 for the creation of mobile, human-centered intelligent multimodal interfaces. Three "lanes" in the road identify three areas of research and development, including empirical and data driven models of multimodality, advanced methods for multimodal communication and toolkits for multimodal systems. The end of the road maps indicate the outcome in 2005, specifically multimodal corpora, computational models, and interface toolkits. Of course there are a variety of interim outcomes along with way. For multimodal corpora this includes annotated corpora of human and natural phenomena (e.g., surveillance, meeting, or broadcast news video) as well as human-machine interactions. Corpora can be used by systems for training or testing/evaluation purposes. In the methods lane, this includes developments such as multimodal mutual disambiguation, multiparty interaction, and multimodal barge in. With respect to toolkits, developments include markup standards for multimodal phenomena (e.g., for combinations of speech, gesture, and facial expressions), reusable components for multimodal analysis and generation, and tools for universal and mobile multimodal access.

[1] Some slides are available at www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/

Stock, O. and Zancanaro, M. (eds.), Intelligent Information Presentation, 1—6.

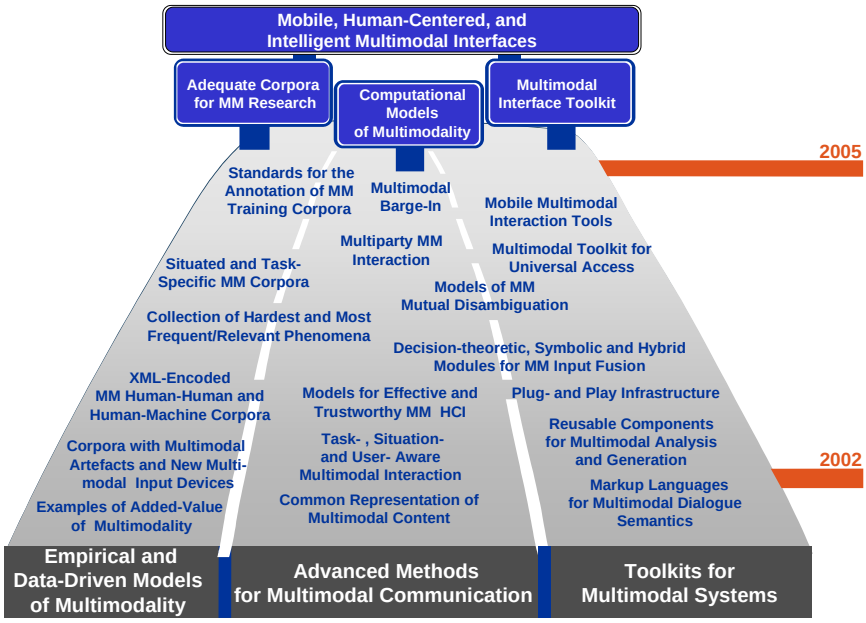

Figure 1. Near Term RoadMap

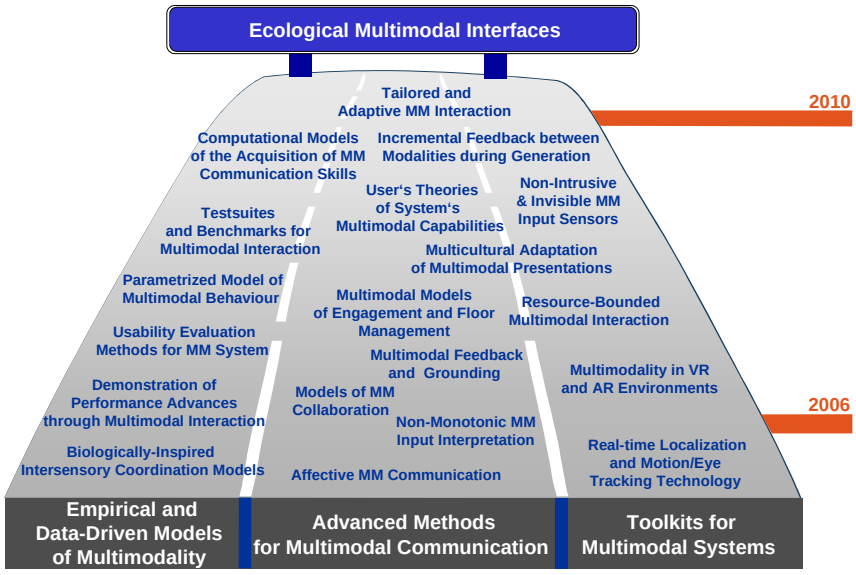

Figure 2. Long Term RoadMap

Figure 2 illustrates the roadmap in the far term, from 2006-2010. Using the same three dimensions as in the short term roadmap, we envision a variety of future outcomes. From 2006 to 2010, in the area of models of multimodality, we envision biologically-inspired intersensory coordination models, test suites and benchmarks, and eventually computational models of the acquisition of multimodal communication skills, among other advancements. Advanced methods will include affective, collaborative, and multicultural multimodal communication. Toolkits will advance from real-time localization and motion/eye tracking, to the incorporation of multimodality into virtual and augmented reality environments, and resource bounded multimodality.

Multimodal Input	Multimodal Interaction	Multimodal Output
● Sensor Technologies	● User Modelling	● Smart Graphics
● Vision	● Cognitive Science	● Design Theory
● Speech & Audio Technology	● Discourse Theory	● Embodied Conversational Agents
● Biometrics	● Ergonomics	● Speech Synthesis

● Machine Learning ● Formal Ontologies ● Pattern Recognition ● Planning

Figure 3. Enabling Technologies

2.0 ABSTRACT ARCHITECTURE

In addition to establishing a roadmap for future research, the Dagstuhl seminar focused on articulating a common reference architecture to consolidate current understanding, facilitate systems description, and to help formulate future systems research. Intelligent multimodal systems require a number of essential functional and technical requirements. Functionally, they need to:

- support modality integration (both fusion of input and design of coordinated output),
- provide situation (User, task, application) appropriate real-time sensing/response (e.g., supporting barge-in, perceptual sensing/feedback),
- represent (modules and data structures) a varying level of granularity manage feedback, both locally and globally
- support incremental processing support incremental development
- be scaleable.

In addition, these functional requirements, there are a number of important system/technical requirements these systems should exhibit, including

- technical means for processing/fusing multimodal input (e.g., parallel processing)
- modular, composable elements and algorithms (possibly distributed processing)
- efficient algorithms and efficient implementations of those
- support for varying time scales, and temporal and spatial resolutions (as well as of course temporal resolution)
- shared (even after partial processing) data structures
- open and extensible protocols for interprocess and intermodule communication.

The group analyzed approximately a dozen architectures of intelligent multimodal systems to modify the base architecture articulated in Maybury and Wahlster (1998) to create the extended and refined architecture shown in Figure 4. The architecture utilizes the definitions of media as a material-centered notion including interactive devices (e.g., keyboard, mouse, microphone) and artifacts (audio, video, text, graphics), mode as human-centered perceptual processes (e.g., visual, auditory, tactile), and code as the formal languages that specific the elements, syntax, semantics, pragmatics and so on that govern the use of media and modes. As can be seen in the figure, this abstract architecture includes functionality for media input processing and media output rendering as well as deeper media/mode analysis and synthesis, which would draw upon at least underlying models of media and modes (language, graphics, gesture). Following analysis, multimodal input would need to be fused and then interpreted within the current state of the discourse, context (time, space, task, domain and so on) and user model including such functions as cross-modal mutual disambiguation. Once the intention of the user (in an interactive setting) had been recognized, the system might interact with the backend application (possibly initiating or terminating sessions, requesting and integrating information or responding to application requests). Finally the system might plan a response to the user, which in turn might require the design of a multimodal presentation (including Content selection, media design, allocation, coordination, layout) which would then need to be synthesized and rendered on specific media for the user. Underlying all the modules in this architecture are mechanisms for representation and inference on a broad range of models. These include models of the user (identity, capabilities, beliefs, and intentions) and other agents (e.g., system, software agents, intermediaries), a model of the discourse (to help track attention and information about interlocuter turns and also detect and correct errors), context (e.g., physical/spatial and temporal state), domain, task, applications and, of course, the media and modalities (their properties and any associated codes). The underlying infrastructure needs to maintain the states and histories of these models, which might be shared with many of the processes shown in the abstract architecture.

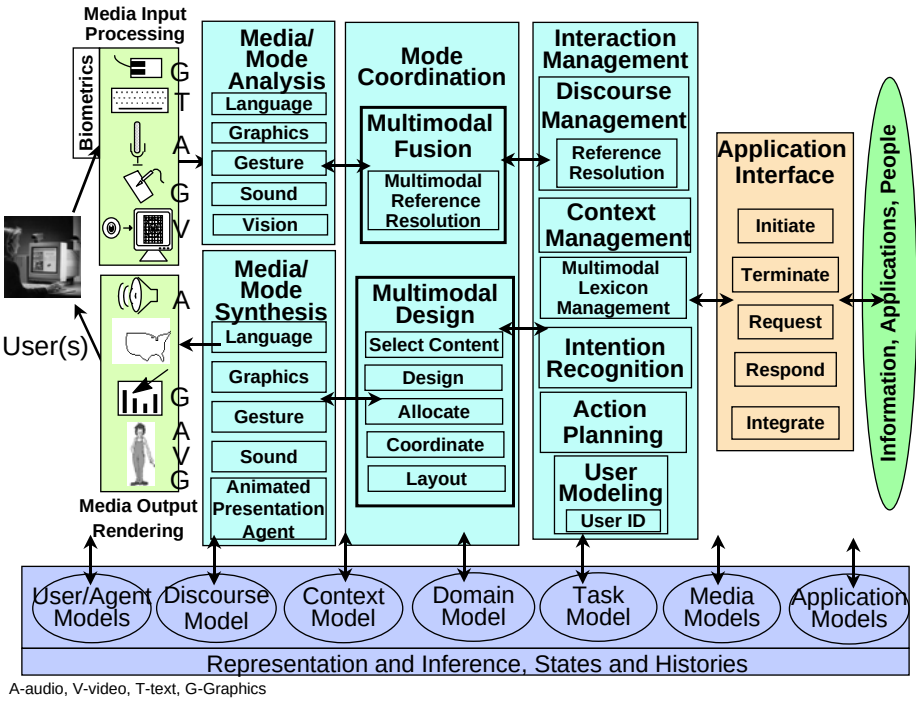

Figure 4. Architecture

3.0 SEMANTICS

Figure 4 and many of the arrows indicating interfunction communication will rely upon an enabling syntax, semantics and pragmatics. A multimodal meaning representation plays central stage in such a system, supporting both interpretation and generation processes. In particular, a multimodal meaning representation should support the fusion and coordination of multiple input- or output modalities at a semantic level, representing the combined and integrated semantic contributions from the different modalities. The interpretation of a multimodal input, such as a spoken utterance combined with a gesture and a certain facial expression, will often have stages of modality-specific processing, resulting in representations of the semantic content of the interactive behaviour in each of the modalities involved. Other stages of interpretation combine and integrate these representations, and take contextual information into account, such as information from the domain model, the discourse model or the user model. A multimodal meaning representation language should support each of these stages of interpretation, as well as the various stages of multimodal output generation and in that sense support *incremental* construction and

processing of semantic content. To make incremental processing feasible, where possible the representations of various types of input and output should be uniform, in the sense of using the same kinds of building blocks and the same ways in which complex structures can be composed of these building blocks. Moreover, to support the representation of partial and intermediate results of semantic interpretation, the framework should allow meaning representations which are *underspecified* in various ways and capture unresolved ambiguities.

When we are considering inputs and outputs from a semantic point of view, the representation of lower-level modality-specific aspects of interactive behaviour, like syntactic linguistic information or morphological properties of gestures is not a primary aim, but some such information may percolate as features associated with a meaning representation, especially at intermediate stages of interpretation, where their relevance for semantic interpretation may not have been fully exploited. At the other end of interpretation, where understanding is rooted in domain models and ontologies, a multimodal meaning representation language should support the connection with frameworks for defining ontologies and specifying domain models, such as OIL and DAML.

The main objective of defining multimodal meaning representations is to provide a fundamental interface format to represent a system's understanding of multimodal user inputs, and to represent meanings that the system will express as multimodal outputs to the user. This means that the first and foremost basic requirement of a semantic representation framework is that (a) it should be expressive enough to correctly represent the meanings of multimodal messages, and (b) that the representation structures themselves have a formal semantics, i.e., their definition should provide a rigorous basis for reasoning (whether deductive, statistical, in the form of plan operators, or otherwise).

In order to delineate the task of formulating objectives, constraints and components of multimodal meaning representation, the group adopted a working definition of meaning in multimodal interaction as the specification of how the interpretation of a multimodal input by an understanding system should change the system's information state (in a broad sense of the term, including domain model, discourse model, user model, task model; cf. Figure 4). While formulated with reference to input interpretation only, this definition can also be related to the generation of multimodal outputs by assuming that an output is generated by the system in order to have an effect on the user through the interpretation of that output by the user. (The generation of appropriate outputs thus depends on the system having an adequate model of what its outputs may mean to the user – which is exactly as it should be.)

An additional objective in defining a well-defined representational framework for multimodal dialogue acts is to allow the specification and comparison of existing application-specific representations (e.g. the M3L representation used in the SmartKom project) and the definition of new ones, while ensuring a level of interoperability between these. Finally, the specification of a multimodal meaning

representation should also form a basis for the definition of annotation schemes of multimodal semantic content. Since the design of multimodal human-computer systems is an area in which new research results and new technologies may bring new challenges and new approaches for the representation of multimodal meanings, the representational framework should be o*pen* to contributions from different theories and approaches, and should be *extensible*, inviting the use of alternative methods for designing representation schemas, like XML.

As a first step in the direction of defining a generic multimodal semantic representation form, we have to establish some basic concepts and corresponding terminology. First, the action-based concept of meaning mentioned above, applicable to multimodal inputs in an interactive situation, means that the meaning of a multimodal `utterance' has two components: one that is often called `propositional' or `referential' and that is concerned with the entities that the utterance refers to and with their properties and relations, and a `functional' component that expresses a speaker's intention in producing the utterance: what effects does he want to achieve with this utterance (using `speaker' in a broad, multimodal sense here)? This distinction is familiar from speech act theory, where the two components are called `propositional content' and `illocutionary force', and is also prevalent in other theories of language-based communication; it is sometimes viewed as drawing a line between semantics and pragmatics. In the analysis of multimodal interaction it is particularly important to pay attention to these two aspects of meaning, since different modalities often contribute to each aspect in different ways; for instance, in spoken interaction the referential and propositional aspects of meaning are often expressed verbally, while gestures and facial expression contribute primarily to functional aspects. The term `multimodal content' should not be confused with `propositional content', and should not make us forget that multimodal messages have meanings with functional aspects that are equally important as their propositional and referential aspects. In this document we use `multimodal content' as synonymous with `multimodal meaning', including functional aspects, and we use `semantic representation' as synonymous with `representation of meaning'.

A convenient term that has become popular in the literature on (multimodal) human-computer dialogue is *dialogue act*. Though mostly used in an informal, intuitive way, or as a variant of `speech act', the term also has a formal definition in terms of the effects that a `speaker' intends to achieve through its understanding by the addressee (see Bunt, 2000), which makes it suitably precise for use in the analysis of the meaning of multimodal inputs and outputs. Without further going into definitions here, we will use the term `dialogue act' in the rest of this document. Definitions of other useful concepts can be found in Romary (2002).

As a second methodological step, we propose to distinguish the following three basic types of ingredients that would seem to go into any multimodal meaning representation framework.

1. *Basic components*: the basic constructs for building representations of the meanings of multimodal dialogue acts: types of building blocks and ways to connect them.
2. *General mechanisms*: representation techniques like substructure labeling and structure sharing, that make the representations more powerful, more flexible and more compact.
3. *Contextual data categories and values*: types of administrative (meta-) data that do not, strictly speaking, contribute to the meanings of semantic representations, but that may nonetheless be relevant for their processing.

Initially, at least the following basic components will be needed for representing the general organization of any semantic structure:

1. Temporal structures (`*events*'), to represent communicative events, like spoken utterances (input or output dialogue acts) and gestures, and semantic events, such as states and events representing meanings of verbs.
2. Referential structures (`*participants'*), to represent for instance the speaker of an input utterance, the addressee of a system output dialogue act, or the individuals and objects participating in a semantic event.
3. *Restrictions* on temporal and referential structures, to represent for instance the type(s) of dialogue, act associated with an utterance, a gesture type, assigned to a gesture token, or the denotations of linguistic modifiers.
4. Dependency structures, representing *semantic relations* between temporal and/or referential structures, such as participant roles (like SPEAKER, ADDRESSEE, AGENT, THEME, SOURCE, …), discourse/rhetorical relations and temporal relations.

To this, other components will have to be added, for instance to represent quantified entities, logically complex restrictions, and propositional attitudes.

General mechanisms like substructure labeling and substructure sharing are important to make meaning representations suitable for partial and underspecified meanings, to give representations a more manageable form, and to relate them to external sources of information. For example, allowing labels instead of the substructures that they label in argument positions opens the possibility of argument underspecification by means of label variables. Structure sharing makes it possible to represent that a certain part of the representation plays more than one role, e.g. a participant may be both the speaker of an utterance and the performer of a gesture, as well as the agent in a verbally expressed semantic event.

Finally, meaning representations will need to be annotated with general contextual information, both globally and also at the level of subexpressions, to capture information which is not found inside the elements of interactive behaviour, but which is potentially relevant for their interpretation and generation, such as environment data (e.g., time stamps and spatial information), processing information (e.g., which module has produced this representation; what is its level of

confidence), and interactional information (who is the speaker; what other addressees are there, etc.).

```
<semRep id="rep1">
  <event id="e0">
    <evtCat>utterance</evtCat>
    <speaker target="Peter"/>
    <adressee target="System"/>
    <alt>
     <dialAct cert="0.8">
        Order</dialAct>
     <dialAct cert="0.3">
        Inform</dialAct>
    </alt>
  <event id="e1">
    <tense>present</tense>
    <evtType>wanttogo</evtType>
    …
  </event>
  <participant id="x">
    <lex>I</lex>
    <synCat><Pronoun</synCat>
    <num>sing</num>
    <pers>first</num>
    …
  </participant>
  <participant id="y">
    <lex>Nancy</lex>
    <synCat>ProperNoun</synCat>
    <pers>third</num>
    …
  </participant>
  <participant id="z">
    <lex>Stuttgart</lex>
    <synCat>ProperNoun</synCat>
    <pers>third</num>
    …
  </participant>
  <relation source="x" target="e1">
    <role>agent</role>
  </relation>
  <relation source="y" target="e1">
    <role>source</role>
  </relation>
  <relation source="y" target="e1">
    <role>goal</role>
  </relation>
  <event id="e2">
    <evtCat>gesture</evtCat>
    <gesturer target="x" >
    <when>2001-11-1:tttt/when>
    <gestType>designation</gestType>
    <graphContext target="ctxt23"/>
     …
  </event>
      …
</semRep>
```

Figure 5. Example of Multimodal Semantic Representation

We illustrate the possible combination of basic components, general mechanisms, and contextual data into a multimodal meaning representation and exemplify the general methodology that we suggested here, by taking up a sample semantic representation derived from an initial example expressed in the ULF+ format (ULF+ is a slightly updated version of a representation language that was developed successively in the PLUS dialogue project, see Geurts and Rentier, 1993, and the multimodal DENK project; see Bunt et al., 1998; Kievit, 1998). Figure 5 shows a fragment of a (simplified) possible XML-style multimodal meaning representation of the utterance "I want to go from here to there" uttered by a speaker named Peter who points at locations on a map while speaking. The top-level element <semRep> corresponds to the multimodal event, consisting of the event of the spoken utterance "e0" and gesture events like "e2". The <event> construct is used both to represent these events and to represent the linguistic content ("e1"); the <participant> element is used to represent the various entities involved in the events. Events and participants are related by means of <relation> elements (with `source` and `target` attributes pointing to the corresponding arguments of the relation. Note the use of the `<alt>` construct to represent an unresolved ambiguity in the interpretation of the utterance as a certain kind of dialogue act and the `cert` attribute for representing the corresponding confidence.

4.0 TOOLS

On the roadmap in Section 1 we see a whole lane dedicated to "Empirical and Data-driven Models of Multimodality" with a big sign ahead saying: "Adequate Corpora for MM Research". Such "adequate corpora" are the fuel for those systems that want to rush down this lane, they are fundamental for the design of representations (Section 3), they have an impact on architectures (user interface, multimodal fusion/integration, see Section 2), and serve as training/test material in data-driven systems. At the Dagstuhl Seminar we explored needs and existing resources for multimodal corpora and annotation tools. We started out by clarifying what we understood to be a multimodal corpus: what kind of data is most fundamental now and what additional needs will come up in the future. Current resources for multimodal research were assessed, including repositories for corpora and tools, institutes and initiatives for data collection. For researchers about to start a data collection effort we wondered about how they could locate technical guidelines, tools and coding schemes. To get an immediate picture on demands and needs we conducted a questionnaire[2] study at the Dagstuhl Seminar, collecting replies from 28 participating researchers coming from a variety of different backgrounds. The analysis revealed a significant lack in the reuse of resources (corpora and tools) due to non-standardization, at the same time establishing a strong interest in a global coordination of resources. This led us to ask what organisational infrastructure

[2] The format of this questionnaire was modified from the questionnaire format elaborated in the ISLE project. The questionnaires are available at http://www.limsi.fr/Individu/martin/questionnairesDagstuhl/

(European and American funding programmes, networks, conferences, workshops, journals, and institutes) there is to support worldwide cooperation. We identified shortcomings and outlined issues for coming conferences and workshops. They are addressed on a workshop at the LREC 2002 conference.

What is a Multimodal Corpus? A multimodal corpus contains primary data (audio/video/text files) and encodings on different layers, descriptive and interpretative ones, for each modality. For sound, a corpus should have a number of standard encodings for human language (transcription, part-of-speech, syntax, co-reference, rhetorical relations, dialogue acts etc.), possibly conforming to a standard like MPEG-7[3]. Besides language, music and environmental noise must be dealt with in descriptive and interpretative terms. For vision, research is currently investigating description of nonverbal communication through the human body, usually focussing on the face (see FACS[4]) or hands/arms (gestures). Posture has recently been picked up. For the future we will need description/meaning encodings of general visual environments, be they static (museum) or moving (car driving scenario, flight simulator). One challenge here is to remodel annotation tools to cope with spatial and spatio-temporal encodings since virtually all current tools base their encodings on time. New to research is the haptic modality: pressure on hands, feet or back (force feedback), even texture can be conveyed (cf. PHANTOM[5]). Encoding aside, we need to ask: how does the primary data look like? Ideally, we would have haptic data as another track in a video file, just like sound is integrated. The same applies for biometric data: heartrate, eye dilation, skin sensitivity, breathing cycles etc. – a modality where communication solely works in the direction human to computer. For the virtual reality community, we will have to look at smell/taste, sense of balance and thermal perception in terms of primary data and encoding formats. More general concepts like mirror behavior, synchronized and repeated behavior, distance or touch, which encompass more than one subject, probably crossing modalities, need to be tackled. It is here where one is forced to draw a line to higher level behavioral and social units which are used in psychology, anthropology etc. They are beyond the scope of standardization efforts for now.

Coding Schemes and Tools As Bird and Liberman (2001) lucidly demonstrated, tools are tailored to a specific annotation framework. If we extract the logical level of an annotation framework and build a core engine with a programmers' interface (API), we could reuse this core for a large number of applications (annotation, visualization, query, analysis etc., see Figure 6). Two such annotation frameworks are the track-based framework (TBF) and the Annotation Graph framework (AGF). The latter is more general than the former, but not all applications may need the added power of the AGF. Isolating and opening up the locial level is a leap forward in standardization but in the near future a single, standard core engine used by all researchers is unlikely to come up. On the other hand, a number of coding schemes

[3] http://mpeg.telecomitalialab.com
[4] Facial Action Coding System (Ekman, Friesen, 1978)
[5] http://www.sensable.com/haptics/haptics.html

have undergone some standardization, especially in the linguistic community (syntax trees, parts-of-speech, dialogue acts etc.), so one may ask whether a unified coding scheme specification language as under development by the ISO/TC 37/SC 4 committee (Ide, Romary 2001) would help in reusing such schemes and make annotated data accessible to applications that work on those schemes (parsers, speech recognizers, generation planners etc.). We envision a standardized specification language for schemes which is independent of a specific framework, thus being independent of a specific tool. It should allow modular extensibility so that new kinds/classes of information can be added easily from outside sources (see Figure 6). Central repositories could collect, store and distribute sets of standard taxonomies (part-of-speech, syntax trees, gesture categories, emblem lexicons etc.) that can be downloaded and readily integrated in a plug-and-play fashion into research schemes. A question in a similar direction is how close an coding scheme's representation can come to a multimodal system's representation language (see Section 3), so the same representation would be used in empricial research as well as in the running system. Such a close match would shorten development time considerably. For tools the question of ergonomic design must be emphasized as more and more mass data is collected, making efficiency a central factor. An integrated user interface should not only allow a smooth annotation workflow but also offer complex search options, visual access to coding schemes, and semi-automatic documentation facilities. The technologies of multimodal interfaces that spring from the annotated data will hopefully itself become part of annotation tools for intuitive and therefore efficient encoding of data, using color, sound, touch etc. Standardized metaphors for coding tool interfaces should develop in the process. Bootstrapping techniques can increase efficiency, especially where standard taxonomies are used (POS, syntax etc.), but may also bias the coder to a certain degree. Multi-coder annotation should be supported by offering update/merge functions (versioning), concurrent coding and reliability checks. To help interpretation of data, more effort is needed in standardization of analysis methods, be they simple descriptive statistics or automatic extraction of meaningful entities, and in the development of evaluation metrics.

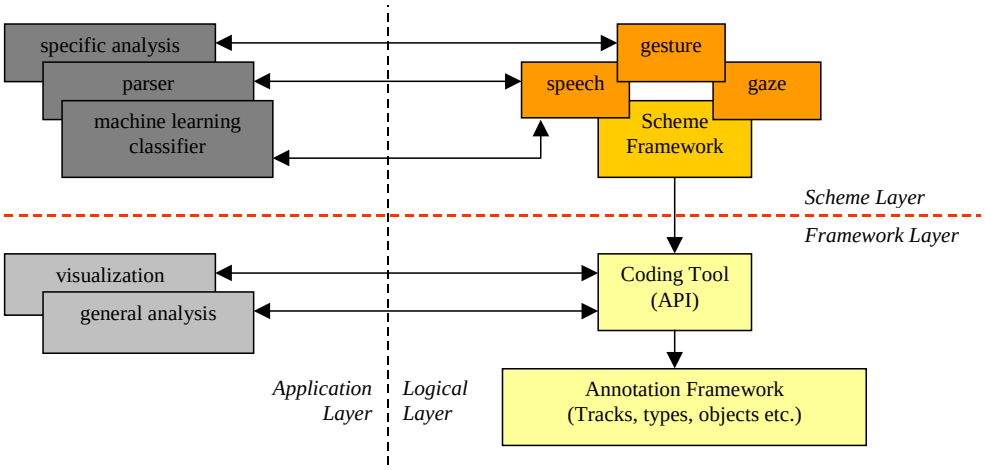

Figure 6. Refined view on the separation between application and logical layer (left/right). In the lower half applications are based on the generic annotation framework, whereas in the upper half they are based on specific coding schemes.

Organisations (Existing Corpora and Tools) We assembled a meta-survey of sources which provide each detailed surveys of corpora, tools and other resources. Survey activities on mulimodal resources are undertaken in the ISLE[6] project (formerly EAGLES), especially the NIMM[7] subgroup (Knudsen et al. 2002a, 2002b), in the TalkBank[8] project and at MITRE (Bigbee et al. 2001). Extensive documentation of corpora and tools is available in papers and websites. General data collection and standardization initiatives in the US are NIST[9] and LDC[10], in Europe ELRA[11], ELDA[12], AHDS[13] (UK), in Japan COCOSDA[14]. Initiatives to build

[6] International Standards for Language Engineering, http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
[7] Natural Interactivity and Multimodality, http://isle.nis.sdu.dk
[8] http://www.talkbank.org
[9] National Institute of Standards and Technology, http://www.nist.gov
[10] Linguistic Data Consortium, http://www.ldc.upenn.edu
[11] European Language Resources Association, http://www.elda.fr
[12] The European Language Resources Distribution Agency, http://www.elda.fr, is ELRA's operational body.
[13] Arts Humanities Data Service

standard tools are ATLAS[15] in the US and NITE[16] in Europe (successor of MATE). ELRA fosters the founding of central national agencies for the collection of native language corpora, and organizes the International Conference on Linguistic Resources and Evaluation[17] (LREC).

Data Collection One crucial issue for data collection is cost reduction and hence knowledge of best practice and potential reusability. Reusability requires standard procedures and formats for collecting primary data (video/audio) and standards for transcription (best practice and coding scheme). Since, due to the diverse needs and goals of different projects, this is impossible to achieve on a large scale, one can retreat to a situation where all corpora are at least easy to locate and browse. A corpus must be identifiable as relevant to your own project by means of *meta-data* that informs in a concise and standardized way about technical setup, file formats and schemes used. Such considerations have led to the founding of OLAC[18] (Bird, Simons, 2001), based on a standard resource description model: the Dublin Core Metadata Set[19] (DCMS). The ISLE MetaData Initiative[20] (IMDI) is also working on meta-data, specifically for multimedia/multimodal corpora, including a mapping to/from OLAC elements. Apart from meta-data, guidelines for best practice are urgently needed. These should include advise on coding scheme design and coding procedures but also on fund-raising and legal issues (ethical and commercial – usually country-specific). The Oxford Text Archive (OTA) is working on an online best practice guide, commissioned by the AHDS. ISLE, Talkbank and LDC also provide best practice guidelines.

Dagstuhl Questionaire Corpora and tools were the main issues of the Dagstuhl questionaire, what was there and what was needed. We collected 28 completed questionaires. The subjects came from 24 different institutes. We will only give the most interesting results. So we found that the dominating modality studied is still speech, closely followed by gesture. Facial expression, posture and gaze were less frequent. 20% of the data is still in analogue format showing that some institutes still use VCR technology. As for the tool situation, 38% use no tool for annotation (just a text editor) and 43% developed their own tool. Clearly dominating scenarios were tourism and navigation, the dominating language was English, followed by German and Japanese, then French and Italian. Of the 28 replies, 21 stated that they were going to collect/code data in the near future. We concluded that on one hand, there is an overlap in research aims (examined modality, language and scenario), on the

[14] International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques, http://www2.slt.atr.co.jp/cocosda
[15] Architecture and Tools for Linguistic Analysis Systems, http://www.nist.gov/speech/atlas
[16] Natural Interactivity Tools Engineering, http://nite.nis.sdu.dk
[17] http://www.lrec-conf.org
[18] Open Language Archives Community, http://www.language-archives.org
[19] http://dublincore.org
[20] http://www.mpi.nl/ISLE/

other, there is a severe lack of resource coordination. Since most participants are about to collect possibly large corpora soon, this questionaire is yet another reason to call for initiatives, workshops and conferences where resources are made available and standardization issues are advanced.

We suggest that small but concrete case studies be collected and exchanged, exemplifying the different approaches to annotation (schemes and tools). Not only would research results become clearer by giving away the original data, we would also have a more competitive situation where schemes as well as tools are used, compared and possibly enhanced. We also suggest to find sample cases which clearly demonstrate the potentials and benefits of multimodal systems.

As a first step, we planned a hands-on annotation exercise which is now part of the workshop on "Multimodal Resources and Multimodal Systems Evaluation"[21], organized by Mark Maybury and Jean-Claude Martin, at LREC 2002,. Major issues are discussion of corpora, schemes, evaluation metrics and annotation tools. We expect researchers from many heterogeneous fields to come to this and similar future workshops on multimodal corpora, from social as well as computational sciences, sharing expertise and resources, and thus making some progress on the road ahead.

5.0 FUTURE RESEARCH

Many outstanding research problems must be solved to realize automatically created user tailored answers to questions. Important issues include:

1. Multimodal ontologies, including representations of time and space for processing multimodal inputs and outputs.
2. New devices for cross modal mutual disambiguation
3. New models for multimodal interaction, for managing context, discourse, user, and application interactions.

We also need a series of community activities and events to ensure continued scientific progress. This includes:

1. Community agreement of standards and methodologies for annotation and markup.
2. Empirical data and user studies to understand fundamental properties of human processing of multimodal (audio, imagery, tactile) input and ouput, and the effects on and limitations of human memory and retention.
3. Evaluation tasks, benchmarks and methods.
4. Regular exchange of knowledge

[21] 1 June 2002, Las Palmas, Gran Canaria, see http://www.lrec-conf.org/lrec2002/index.html

5. Shared or at least interoperable tools to ensure progress.

6.0. AFFILIATIONS

[2]German Research Center for AI
DFKI
Stuhlsatzenhausweg 3
66123 Saarbruecken, Germany
{ wahlster, kipp?}@dfki.de
www.dfki.de/~wahlster, ~kipp

[3]Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
Maybury@mitre.org

[1]Linguistics and Computer Science
Tilburg University
P.O. Box 90153
5000 LE Tilburg, the Netherlands
Harry.Bunt@kub.nl

7.0 CONCLUSION

Given the overload of information and knowledge, the growth in heterogeneous computing platforms, and the increasing ubiquity of communications and information access for fixed and mobile users, intelligent interaction may prove to be the most important application in the next decade. Achieving some of the key issues we have outlines in the paper will help out societies reach promising facilities for important sociotechnical objectives such as information access for all, increased task performance and higher quality interactions.

REFERENCES

Bigbee, T. and Loehr, D. and Harper, L. 2001. “Emerging Requirements for Multi-Modal Annotation and Analysis Tools”. In: *Proceedings of Eurospeech*, pages 1533-1536.

Bird, S. and Liberman, M. 2000. “A Formal Framework for Linguistic Annotation”. In: *Speech Communication* **33** (1/2), pages 23-60.

Bird, S. and Simons, G. (2001). “The OLAC Metadata Set and Controlled Vocabularies”. In: Proceedings of the ACL/EACL Workshop on Sharing Tools and Resources, pages 7-18.

Bunt, H. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors. Abduction, Belief and Context in Dialogue. Amsterdam: John Benjamins Publishing Company. (ISBN 90-272-4983-0 (Eur.)/1-55619-794-2 (US)

Bunt, H. and Beun, R.J. editors, 2001. Cooperative Multimodal Communication. Springer Lecture Series in Artificial Intelligence 2155. Berlin: Springer Verlag. (ISBN 3-540-42806-2)

Bunt, H.C., R. Ahn, R.J. Beun, T. Borghuis and C. van Overveld, 1998. Multimodal cooperation with the DENK system. In: H.C. Bunt, R.J. Beun and T. Borghuis, editors Multimodal Human-Computer Communication. Berlin: Springer Verlag.

Ekman, P. and Friesen, W.V. (1978). Facial Action Coding System. Palo Alto, CA: Consulting Psychologists Press.

Geurts, B. and G. Rentier, 1993. Quasi-logical form in PLUS. Internal Report, Esprit Project P5254, A Pragmatics-based Language Understanding System. Tilburg University.

Ide, N. and Romary, L. (2001). "Standards for Language Resources", In: Proceedings of the IRCS Workshop on Linguistic Databases, pages 141-149.

Kievit, L.A., 1998. Context-driven Natural Language Interpretation. Ph.D. Thesis, Tilburg University.

Kita, S. and van Gijn, I. and van der Hulst, H. (1998). "Movement Phases in Signs and Co-speech Gestures and Their Transcription by Human Coders". In: Wachsmuth, I. and Fröhlich, M. (eds.) Gesture and Sign Language in Human-Computer Interaction, pages 23-35.

Knudsen, M. W., Martin, J. C., Berman, S., Bernsen, N. O., Choukri, K., Dybkjær, L., Heid, U., Mapelli, V., Pelachaud, C., and Poggi, I. (2002a). Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources, ISLE Deliverable D8.1, http://isle.nis.sdu.dk.

Knudsen, M. W., Martin, J. C., Bernsen, N. O., Dybkjær, L., Heid, U., Pelachaud, C., Poggi, I., Reithinger, van ElsWijk, G., Wittenburg, P., Llisterri, J. and Ayuso, M. J. M, N., Carletta, J. (2002b). Survey of Multimodal Annotation Schemes and Best Practice, ISLE Deliverable D9.1, http://isle.nis.sdu.dk.

Maybury, M. T. editor. 1993. Intelligent Multimedia Interfaces. AAAI/MIT Press. 405 pp. ISBN 0-262-63150-4 (www.aaai.org:80/Press/Books/Maybury1, mitpress.mit.edu/book-home.tcl?isbn=0262631504)

Maybury, M. T. and Wahlster, W. editors. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Press. (www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-444-8)

Romary, L., 2002. MMIL requirements specification. Project MIAMM – *Multidimensional Information Access using Multiple Modalities.* EU project IST-20000-29487, Deliverable D6.1. LORIA, Nancy.

ACKNOWLEDGEMENTS

The ideas expressed in this document reflect the contributions of all participants at the Daghstuhl Workshop, including the authors and Sharon Oviatt (OGI), Oliviero Stock (IRST), Rainer Malaka (European Media Lab), Elmar Nöth (Univ. of Erlangen), Norbert Reithinger (DFKI), Koiti Hasida (ETL), Noelle Carbonell (Loria), Candy Sidner (MERL), Laurent Romary (UMR Loria), Justine Cassell, (MIT Media Lab) Derek Jacoby(Microsoft Research), Dagmar Schmauks (TU University), Robbert-Jan Beun (Utrecht University), Jean-Claude Martin (Univ. Paris 8), Emiel Krahmer (Tilburg University), Fabio Pianesi (IRST), Elisabeth Andre (Univ Aachen), Thomas Rist (DFKI), Arne Jönnson (Linköping University), Catherine Pelachaud (Rome), Paul Mc Kevitt (Univ. of Ulster, Londonderry), John Lee (Univ. of Edinburgh) and Lisa Harper (MITRE).