In: Wahlster, W. (ed.) (2006): SmartKom: Foundations of Multimodal Dialogue Systems. Cognitive Technologies Series, Heidelberg, Germany: Springer, 2006, pp. 3-27

Dialogue Systems Go Multimodal: The SmartKom Experience

Wolfgang Wahlster

DFKI GmbH, Saarbrücken, Germany
wahlster@dfki.de

Summary. Multimodal dialogue systems exploit one of the major characteristics of human-human interaction: the coordinated use of different modalities. Allowing all of the modalities to refer to and depend upon each other is a key to the richness of multimodal communication. We introduce the notion of symmetric multimodality for dialogue systems in which all input modes (e.g., speech, gesture, facial expression) are also available for output, and vice versa. A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output. We present an overview of the SMARTKOM system that provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent. SMARTKOM represents a new generation of multimodal dialogue systems that deal not only with simple modality integration and synchronization but cover the full spectrum of dialogue phenomena that are associated with symmetric multimodality (including crossmodal references, one-anaphora, and backchannelling). We show that SMARTKOM's plug-and-play architecture supports multiple recognizers for a single modality, e.g., the user's speech signal can be processed by three unimodal recognizers in parallel (speech recognition, emotional prosody, boundary prosody). We detail SMARTKOM's three-tiered representation of multimodal discourse, consisting of a domain layer, a discourse layer, and a modality layer. We discuss the limitations of SMARTKOM and how they are overcome in the follow-up project SmartWeb. In addition, we present the research roadmap for multimodality addressing the key open research questions in this young field. To conclude, we discuss the economic and scientific impact of the SMARTKOM project, which has led to more than 50 patents and 29 spin-off products.

1 The Need for Multimodality

In face-to-face situations, human dialogue is not only based on speech but also on nonverbal communication including gesture, gaze, facial expression, and body posture. Multimodal dialogue systems exploit one of the major characteristics of human-human interaction: the coordinated use of different modalities. The term *modality* refers to the human senses: vision, audition, olfaction, touch, and taste. In addition, human communication is based on socially shared code systems like natural languages, body languages, and pictorial languages with their own syntax, semantics,

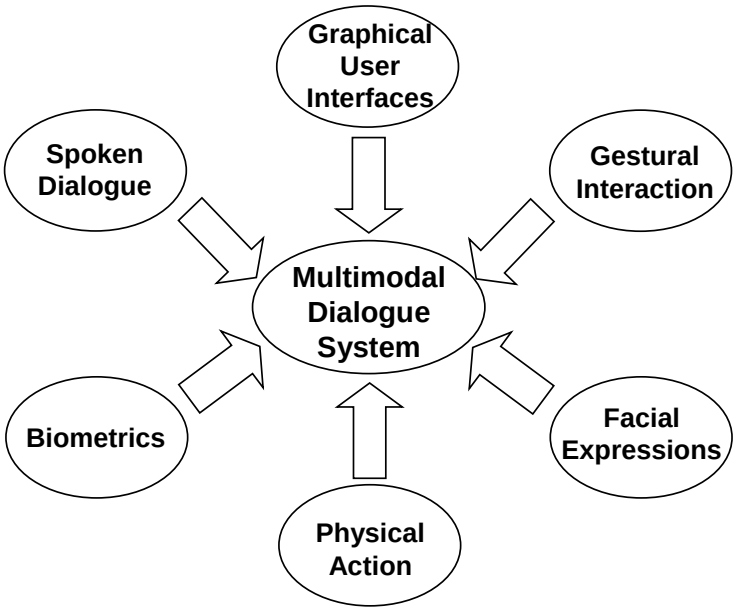

Fig. 1. Merging various dialogue and interaction paradigms into multimodal dialogue systems

and pragmatics. A single semiotic code may be supported by many modalities (Maybury and Wahlster, 1998). For instance, language can be supported visually (i.e., written language), aurally (i.e., spoken language) or tactilely (i.e., Braille script) – in fact, spoken language can have a visual component (e.g., lipreading). Allowing all of the modalities to refer to and depend upon each other is a key to the richness of multimodal communication (see Fig. 1).

Unlike traditional keyboard and mouse interfaces or unimodal spoken dialogue systems, multimodal dialogue systems permit flexible use of input and output modes (Oviatt, 2003). Since there are large individual differences in ability and preference to use different modalities, a multimodal dialogue system permits diverse user groups to exercise control over how they interact with application systems. Especially for mobile tasks, multimodal dialogue systems permit the modality choice and switching that is needed during the changing situational conditions.

With the proliferation of embedded computers in everyday life and the emergence of ambient intelligence, multimodal dialogue systems are no longer only limited to traditional human-computer communication, but become more generally a key component for advanced human-technology interaction and even multimodal human-environment communication (Wahlster and Wasinger, 2006).

Self-service systems, online help systems, Web services, mobile communication devices, remote control systems, smart appliances, and dashboard computers are providing ever more functionality. However, along with greater functionality, the user must also come to terms with the greater complexity and a steeper learning curve.

This complexity is compounded by the sheer proliferation of different systems lacking a standard user interface. The growing emphasis on multimodal dialogue systems is fundamentally inspired by the aim to support natural, flexible, efficient, and powerfully expressive means of human-computer communication that are easy to learn and use. Advances in human language technology and in perceptive user interfaces offer the promise of pervasive access to online information and Web services. The long-term goal of the research in multimodal dialogue systems is to allow the average person to interact with computerized technologies anytime and anywhere without special skills or training, using such common devices as a smartphone or a PDA.

We begin by describing the scientific goals of the SMARTKOM project in Sect. 2, before we introduce the notion of symmetric multimodality in Sect. 3. In Sect. 4, we introduce SMARTKOM as a flexible and adaptive multimodal dialogue shell and show in Sect. 5 that SMARTKOM bridges the full loop from multimodal perception to physical action. SMARTKOM's distributed component architecture, realizing a multiblackboard system, is described in Sect. 6. Then in Sects. 7 and 8, we describe SMARTKOM's methods for multimodal fusion and fission. Section 9 discusses the role of the three-tiered multimodal discourse model in SMARTKOM. Section 10 gives a brief introduction to SmartWeb, the follow-up project to SMARTKOM, which supports open-domain question answering. Open research questions in the young field of multimodal dialogue systems are presented in Sect. 11 in a research roadmap for multimodality. Finally, we discuss the economic and scientific impact of the SMARTKOM project in Sect. 12.

2 SmartKom: A Massive Approach to Multimodality

In this book, we present the theoretical and practical foundations of multimodal dialogue systems using the results of our large-scale project SMARTKOM as the background of our discussion. Our SMARTKOM system (http://www.smartkom.org) is designed to support a wide range of collaborative and multimodal help dialogues that allow users to intuitively and efficiently access the functionalities needed for their task. The application of the SMARTKOM technology is especially motivated in non-desktop scenarios, such as smart rooms, kiosks, or mobile environments. SMARTKOM features the situated understanding of possibly imprecise, ambiguous or incomplete multimodal input and the generation of coordinated, cohesive, and coherent multimodal presentations. SMARTKOM's interaction management is based on representing, reasoning, and exploiting models of the user, domain, task, context, and modalities. The system is capable of real-time dialogue processing, including flexible multimodal turn-taking, backchannelling, and metacommunicative interaction. The four major scientific goals of SMARTKOM were to:

- explore and design new symbolic and statistical methods for the seamless fusion and mutual disambiguation of multimodal input on semantic and pragmatic levels
- generalize advanced discourse models for spoken dialogue systems so that they can capture a broad spectrum of multimodal discourse phenomena

- explore and design new constraint-based and plan-based methods for multimodal fission and adaptive presentation layout
- integrate all these multimodal capabilities in a reusable, efficient and robust dialogue shell that guarantees flexible configuration, domain independence and plug-and-play functionality

3 Towards Symmetric Multimodality

SMARTKOM provides *full symmetric multimodality* in a mixed-initiative dialogue system. Symmetric multimodality means that all input modes (speech, gesture, facial expression) are also available for output, and vice versa. A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output.

In this sense, SMARTKOM's modality fission component provides the inverse functionality of its modality fusion component, since it maps a communicative intention of the system onto a coordinated multimodal presentation (Wahlster, 2002). SMARTKOM provides an anthropomorphic and affective user interface through an embodied conversational agent called Smartakus. This life-like character uses coordinated speech, gesture and facial expression for its dialogue contributions.

Thus, SMARTKOM supports face-to-face dialogic interaction between two agents that share a common visual environment: the human user and Smartakus, an autonomous embodied conversational agent. The "i"-shape of Smartakus is analogous to that used for *information* kiosks (see Fig. 2). Smartakus is modeled in 3D Studio Max. It is a self-animated interface agent with a large repertoire of gestures, postures and facial expressions. Smartakus uses body language to notify users that it is waiting for their input, that it is listening to them, that is has problems in understanding their input, or that it is trying hard to find an answer to their questions.

Most of the previous multimodal interfaces do not support symmetric multimodality, since they focus either on multimodal fusion (e.g., QuickSet, see Cohen et al. (1997), or MATCH, see Johnston et al. (2002)) or multimodal fission (e.g., WIP, see Wahlster et al. (1993)). But only true multimodal dialogue systems like SMARTKOM create a natural experience for the user in the form of daily human-to-human communication, by allowing both the user and the system to combine the same spectrum of modalities.

SMARTKOM is based on the situated delegation-oriented dialogue paradigm (SDDP, see Fig. 2): The user delegates a task to a virtual communication assistant (Wahlster et al., 2001). This cannot however be done in a simple command-and-control style for more complex tasks. Instead, a collaborative dialogue between the user and the agent elaborates the specification of the delegated task and possible plans of the agent to achieve the user's intentional goal. The user delegates a task to Smartakus and helps the agent, where necessary, in the execution of the task. Smartakus accesses various digital services and appliances on behalf of the user, collates the results, and presents them to the user.

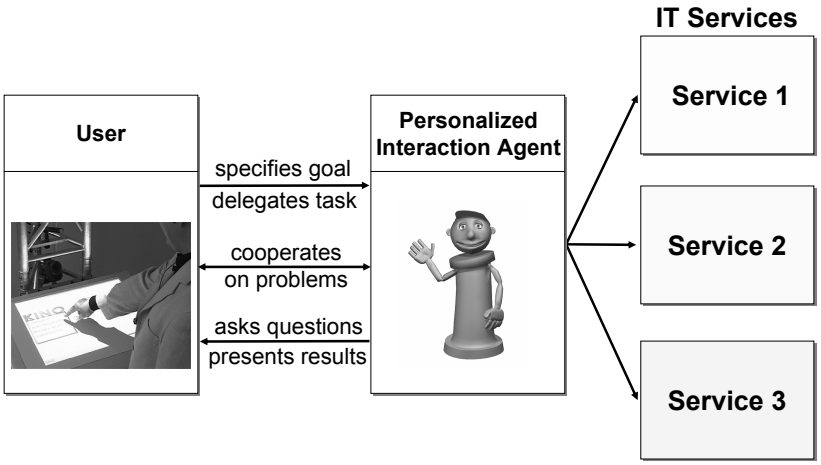

Fig. 2. SMARTKOM's SDDP interaction metaphor

SMARTKOM represents a new generation of multimodal dialogue systems that deal not only with simple modality integration and synchronization but cover the full spectrum of dialogue phenomena that are associated with symmetric multimodality. One of the technical goals of our research in the SMARTKOM project was to address the following important discourse phenomena that arise in multimodal dialogues:

- mutual disambiguation of modalities
- multimodal deixis resolution and generation
- crossmodal reference resolution and generation
- multimodal anaphora resolution and generation
- multimodal ellipsis resolution and generation
- multimodal turn-taking and backchannelling

Symmetric multimodality is a prerequisite for a principled study of these discourse phenomena.

4 Towards a Flexible and Adaptive Shell for Multimodal Dialogues

SMARTKOM was designed with a clear focus on flexibility, as a transmutable system that can engage in many different types of tasks in different usage contexts. The same software architecture and components are used in various roles that Smartakus can play in the following three fully operational experimental application scenarios (see Fig. 3):

- a *communication companion* that helps with phone, fax, email, and authentication tasks

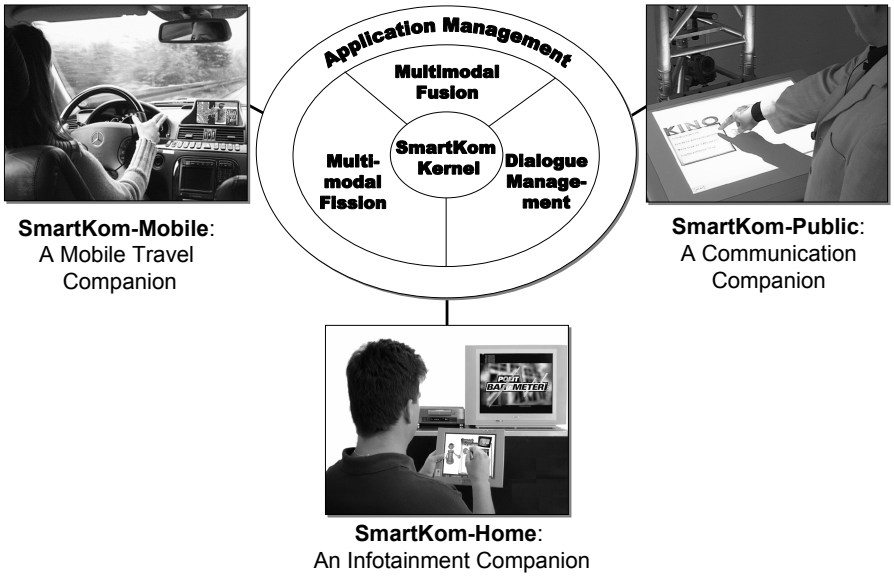

Fig. 3. The three application scenarios of SMARTKOM

- an *infotainment companion* that helps to select media content and to operate various TV appliances (using a tablet computer as a mobile client)
- a *mobile travel companion* that helps with navigation and point-of-interest information retrieval in location-based services (using a PDA as a mobile client)

Currently, the user can delegate 43 types of complex tasks to Smartakus in multimodal dialogues. The SMARTKOM architecture supports not only simple multimodal command-and-control interfaces, but also coherent and cooperative dialogues with mixed initiative and a synergistic use of multiple modalities. SMARTKOM's plug-and-play architecture supports easy addition of new application services.

Figure 4 shows a three-camera configuration of SMARTKOM that can be used as a multimodal communication kiosk for airports, train stations, or other public places where people may seek information on facilities such as hotels, restaurants, and movie theatres. Users can also access their personalized Web services. The user's speech input is captured with a directional microphone. The user facial expressions of emotion are captured with a CCD camera and his gestures are tracked with an infrared camera. A video projector is used for the projection of SMARTKOM's graphical output onto a horizontal surface. Two speakers under the projection surface provide the speech output of the life-like character. An additional camera that can automatically tilt and pan is used to capture images of documents or 3D objects that the user would like to include in multimedia messages composed with the help of SMARTKOM.

As a resource-adaptive multimodal system, the SMARTKOM architecture supports a flexible embodiment of the life-like character that is used as a conversational

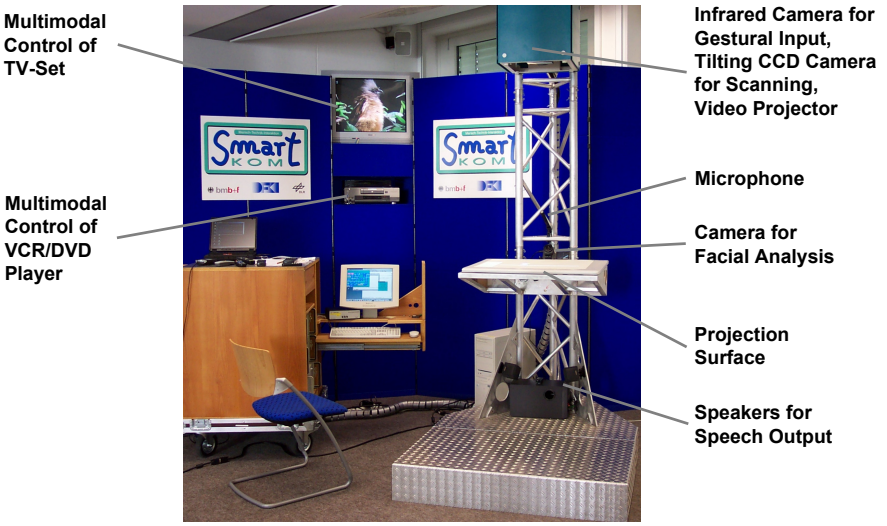

Fig. 4. Multimodal input and output devices of SMARTKOM-Public

partner in multimodal dialogue. The Smartakus agent is visualized either simply as a talking head together with an animated hand, when screen space is scarce, or as a full-body character, when enough screen space is available (see Fig. 2). Thus, Smartakus is embodied on a PDA differently than on a tablet computer or on the large top-projected screen used in the public information kiosk.

5 Perception and Action Under Multimodal Conditions

SMARTKOM bridges the full loop from multimodal perception to physical action. Since the multimodal interaction with Smartakus covers both communicative and physical acts, the mutual understanding of the user and the system can often be validated by checking whether the user and the system "do the right thing" for completing the task at hand.

In a multimodal dialogue about the TV program, the user may browse a TV show database, create a personalized TV listing, and finally ask Smartakus to switch on the TV and tune to a specific program. Smartakus can also carry out more complex actions like programming a VCR to record the user's favourite TV show. Moreover, it can scan a document or a 3D object with its camera and then send the captured image to another person as an email attachment. Fig. 5 shows Dr. Johannes Rau, the former German Federal President, using SMARTKOM's multimodal dialogue capabilities to scan the "German Future Award" trophy and send the scanned image via email to a colleague. This example shows that, on the one hand, multimodal dialogue contributions can trigger certain actions of Smartakus. On the other hand, Smartakus

Fig. 5. The former German Federal President e-mailing a scanned image with the help of Smartakus

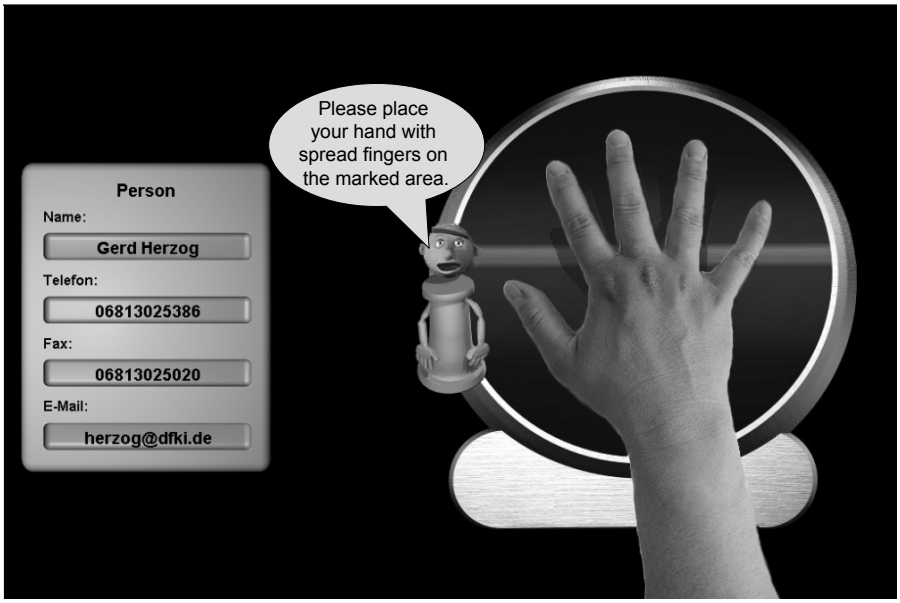

Fig. 6. Interactive biometric authentication by hand contour recognition

may also ask the user to carry out certain physical actions during the multimodal dialogue.

For example, Smartakus will ask the user to place his hand with spread fingers on a virtual scanning device, or to use a write-in field projected on the screen for his signature, when biometric authentication by hand contour recognition or signature verification is requested by a security-critical application. Fig. 6 shows a situation in which Smartakus has found an address book entry for the user, after he has introduced himself by name. Since the address book entry, which is partially visualized

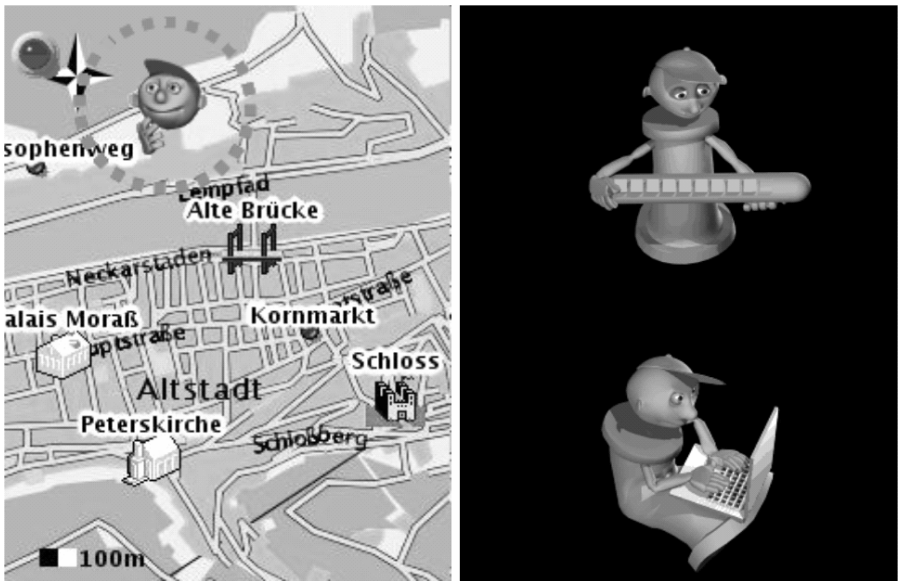

Fig. 7. Adaptive perceptual feedback on the system state

by SMARTKOM on the left part of the display, requests hand contour authentication for this particular user, Smartakus asks the user to place his hand on the marked area of the projected display, so that the hand contour can be scanned by its camera (see Fig. 6).

Since quite complex tasks can be delegated to Smartakus, there may be considerable delays in replying to a request. Our WOZ (Wizard-of-Oz) experiments and user tests with earlier prototypes of SMARTKOM showed clearly that users want simple and fast feedback on the state of the system in such situations. Therefore, a variety of adaptive perceptual feedback mechanisms have been realized in SMARTKOM.

In the upper-left corner of a presentation, SMARTKOM can display a "magic eye" icon, that lights up while the processing of the user's multimodal input is proceeding (see the left part of Fig. 7). "Magic eye" is the common name applied to the green-glow tubes used in 1930s radio equipment to visually assist the listener in tuning a radio station to the point of greatest signal strength. Although SMARTKOM works in real-time, there may be some processing delays caused by corrupted input or complex disambiguation processes.

An animated dashed line (see the left part of Fig. 7) circles the Smartakus character, while the system is engaged in an information retrieval task (e.g., access to maps, EPG (Electronic Program Guide), Web sites). This type of feedback is used when screen space is scarce. When more screen space is available, an animation sequence that shows Smartakus working on a laptop is used for the same kind of feedback. When Smartakus is downloading a large file, it can show a progress bar to indicate to the user how the data transfer is going (see the right part of Fig. 7).

6 A Multiblackboard Platform with Ontology-Based Messaging

SMARTKOM is based on a distributed component architecture, realizing a multiblackboard system. The integration platform is called MULTIPLATFORM (**Mu**ltiple **L**anguage **T**arget **I**ntegration **Plat**form **for M**odules, see Herzog et al. (2003)) and is built on top of open source software. The natural choice to realize an open, flexible and scalable software architecture is that of a distributed system, which is able to integrate heterogeneous software modules implemented in diverse programming languages and running on different operating systems. SMARTKOM includes more than 40 asynchronously running modules coded in four different programming languages: C, C++, Java, and Prolog.

The MULTIPLATFORM testbed includes a message-oriented middleware. The implementation is based on PVM, which stands for *parallel virtual machine*. On top of PVM, a message-based communication framework is implemented based on the so-called publish/subscribe approach. In contrast to unicast routing known from multiagent frameworks, which realize a direct connection between a message sender and a known receiver, MULTIPLATFORM is based on the more efficient multicast addressing scheme. Instead of addressing one or several receivers directly, the sender publishes a notification on a named message queue, so that the message can be forwarded to a list of subscribers. This kind of distributed event notification makes the communication framework very flexible as it focuses on the data to be exchanged and it decouples data producers and data consumers. Compared with point-to-point messaging used in multiagent frameworks like OAA (Martin et al., 1999), the publish/subscribe scheme helps to reduce the number and complexity of interfaces significantly.

GCSI, the **G**alaxy **C**ommunicator **S**oftware **I**nfrastructure (Seneff et al., 1999) architecture is also fundamentally different from our approach. The key component of GCSI is a central hub, which mediates the interaction among various servers that realize different dialogue system components. Within MULTIPLATFORM there exists no such centralized controller component, since this could become a bottleneck for more complex multimodal dialogue architectures.

In order to provide publish/subscribe messaging on top of PVM, we have added another software layer called PCA (**P**ool **C**ommunication **A**rchitecture). In MULTIPLATFORM, the term *data pool* is used to refer to named message queues. Every single pool can be linked with a pool data format specification in order to define admissible message contents. The messaging system is able to transfer arbitrary data contents, and provides excellent performance characteristics (Herzog et al., 2003).

In SMARTKOM, we have developed M3L (**M**ulti**m**odal **M**arkup **L**anguage) as a complete XML language that covers all data interfaces within this complex multimodal dialogue system. Instead of using several quite different XML languages for the various data pools, we aimed at an integrated and coherent language specification, which includes all substructures that may occur on the different pools. In order to make the specification process manageable and to provide a thematic organization, the M3L language definition has been decomposed into about 40 schema specifica-

tions. The basic data flow from user input to system output continuously adds further processing results so that the representational structure will be refined, step-by-step.

The ontology that is used as a foundation for representing domain and application knowledge is coded in the ontology language OIL. Our tool OIL2XSD (Gurevych et al., 2003) transforms an ontology written in OIL (Fensel et al., 2001) into an M3L compatible XML Schema definition. The information structures exchanged via the various blackboards are encoded in M3L. M3L is defined by a set of XML schemas. For example, the word hypothesis graph and the gesture hypothesis graph, the hypotheses about facial expressions, the media fusion results, and the presentation goal are all represented in M3L. M3L is designed for the representation and exchange of complex multimodal content. It provides information about segmentation, synchronization, and the confidence in processing results. For each communication blackboard, XML schemas allow for automatic data and type checking during information exchange. The XML schemas can be viewed as typed feature structures. SMARTKOM uses unification and a new operation called OVERLAY (Alexandersson and Becker, 2003) of typed feature structures encoded in M3L for discourse processing.

Application developers can generate their own multimodal dialogue system by creating knowledge bases with application-specific interfaces, and plugging them into the reusable SMARTKOM shell. It is particularly easy to add or remove modality analyzers or renderers, even dynamically while the system is running. This plug and play of modalities can be used to adjust the system's capability to handle different demands of the users, and the situative context they are currently in. Since SMARTKOM's modality analyzers are independent from the respective device-specific recognizers, the system can switch in real-time, for example, between video-based, pen-based or touch-based gesture recognition. SMARTKOM's architecture, its dialogue backbone, and its fusion and fission modules are reusable across applications, domains, and modalities.

MULTIPLATFORM is running on the SMARTKOM server that consists of 3 dual Xeon 2.8 GHz processors. Each processor uses 1.5 GB of main memory. One processor is running under Windows 2000, and the other two under Linux. The mobile clients (an iPAQ Pocket PC for the mobile travel companion and a Fujitsu Stylistic 3500X webpad for the infotainment companion) are linked to the SMARTKOM server via WaveLAN.

7 Reducing Uncertainty and Ambiguity by Modality Fusion

The analysis of the various input modalities by SMARTKOM is typically plagued by uncertainty and ambiguity. The speech recognition system produces a word hypothesis graph with acoustic scores, stating which word might have been spoken in a certain time frame. The prosody component generates a graph of hypotheses about clause and sentence boundaries with prosodic scores. The gesture analysis component produces a set of scored hypotheses about possible reference objects in the visual context. Finally, the interpretation of facial expressions leads to various scored hypotheses about the emotional state of the user. All the recognizers produce

time-stamped hypotheses, so that the fusion process can consider various temporal constraints. The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results. By fusing symbolic and statistical information derived from the recognition and analysis components for speech, prosody, facial expression and gesture, SMARTKOM can correct various recognition errors of its unimodal input components and thus provide a more robust dialogue than a unimodal system.

In principle, modality fusion can be realized during various processing stages like multimodal signal processing, multimodal parsing, or multimodal semantic processing. In SMARTKOM, we prefer the latter approach, since for the robust interpretation of possibly incomplete and inconsistent multimodal input, more knowledge sources become available on later processing stages. An early integration on the signal level allows no backtracking and reinterpretation, whereas the multimodal parsing approach has to prespecify all varieties of crossmodal references, and is thus unable to cope robustly with unusual or novel uses of multimodality. However, some early fusion is also used in SMARTKOM, since the scored results from a recognizer for emotional prosody (Batliner et al., 2000) are merged with the results of a recognizer for affective facial expression. The classification results are combined in a synergistic fashion, so that a hypothesis about the affective state of the user can be computed.

In SMARTKOM, the user state is used, for example, in the dialogue-processing backbone to check whether the user is satisfied or not with the information provided by Smartakus. It is interesting to note that SMARTKOM's architecture supports multiple recognizers for a single modality. In the current system, prosody is evaluated by one recognizer for clause boundaries and another recognizer for emotional speech. This means that the user's speech signal is processed by three unimodal recognizers in parallel (speech recognition, emotional prosody, boundary prosody).

The time stamps for all recognition results are extremely important since the confidence values for the classification results may depend on the temporal relations between input modalities. For example, experiments in SMARTKOM have shown that the results from recognizing various facial regions (like eye, nose, and mouth area) can be merged to improve recognition results for affective states like anger or joy. However, while the user is speaking, the mouth area does not predict emotions reliably, so that the confidence value of the mouth area recognizer must be decreased. Thus, SMARTKOM's modality fusion is based on adaptive confidence measures that can be dynamically updated depending on the synchronization of input modalities.

One of the fundamental mechanisms implemented in SMARTKOM's modality fusion component is the extended unification of all scored hypothesis graphs and the application of mutual constraints in order to reduce the ambiguity and uncertainty of the combined analysis results. This approach was pioneered in our XTRA system, an early multimodal dialogue system that assisted the user in filling out a tax form with a combination of typed natural language input and pointing gestures (Wahlster, 1991). QuickSet uses a similar approach (Cohen et al., 1997).

In SMARTKOM, the intention recognizer has the task to finally rank the remaining interpretation hypotheses and to select the most likely one, which is then passed

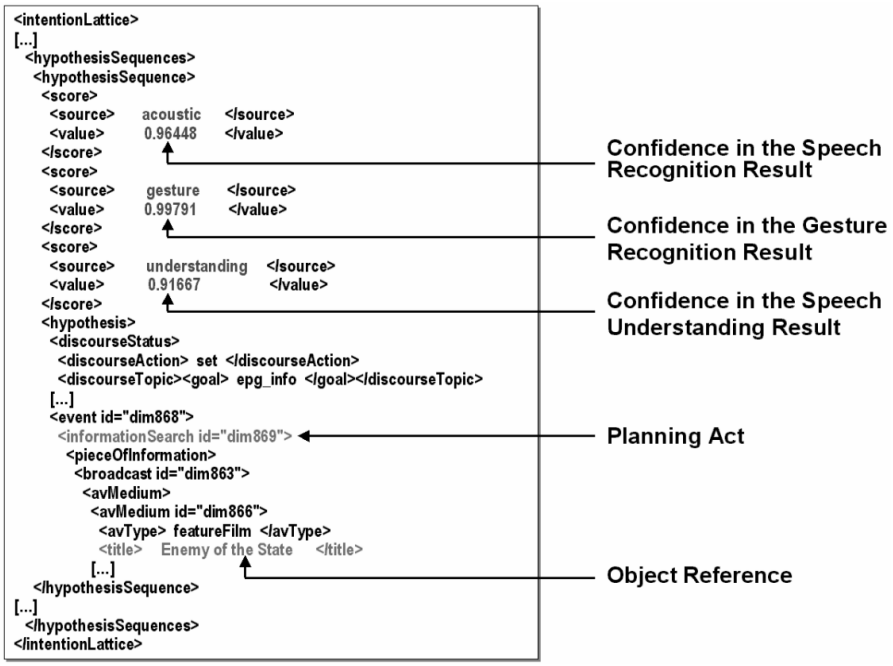

Fig. 8. M3L representation of an intention lattice fragment

on to the action planner. The modality fusion process is augmented by SMARTKOM's multimodal discourse model, so that the final ranking of the intention recognizer becomes highly context sensitive. The discourse component produces an additional score that states how good an interpretation hypothesis fits to the previous discourse (Pfleger et al., 2002). As soon as the modality fusion component finds a referential expression that is not combined with an unambiguous deictic gesture, it sends a request to the discourse component asking for reference resolution. If the resolution succeeds, the discourse component returns a completely instantiated domain object.

Figure 8 shows an excerpt from the intention lattice for the user's input "I would like to know more about this [deictic pointing gesture]". It shows one hypothesis sequence with high scores from speech and gesture recognition. A potential reference object for the deictic gesture (the movie title "Enemy of the State") has been found in the visual context. SMARTKOM assumes that the discourse topic relates to an electronic program guide and the intended action of Smartakus refers to the retrieval of information about a particular broadcast.

8 Plan-Based Modality Fission in SmartKom

In SMARTKOM, modality fission is controlled by a presentation planner. The input to the presentation planner is a presentation goal encoded in M3L as a modality-free representation of the system's intended communicative act. This M3L structure is generated by either an action planner or the dynamic help component, which can initiate clarification subdialogues. The presentation planning process can be adapted to various application scenarios via presentation parameters that encode user preferences (e.g., spoken output is preferred by a car driver), output devices (e.g., size of the display), or the user's native language (e.g., German vs. English). A set of XSLT (Extensible Stylesheet Language Transformations) stylesheets is used to transform the M3L representation of the presentation goal, according to the actual presentation parameter setting. The presentation planner recursively decomposes the presentation goal into primitive presentation tasks using 121 presentation strategies that vary with the discourse context, the user model, and ambient conditions. The presentation planner allocates different output modalities to primitive presentation tasks, and decides whether specific media objects and presentation styles should be used by the media-specific generators for the visual and verbal elements of the multimodal output.

The presentation planner specifies presentation goals for the text generator, the graphics generator, and the animation generator. The animation generator selects appropriate elements from a large catalogue of basic behavioral patterns to synthesize fluid and believable actions of the Smartakus agent. All planned deictic gestures of Smartakus must be synchronized with the graphical display of the corresponding media objects, so that Smartakus points to the intended graphical elements at the right moment. In addition, SMARTKOM's facial animation must be synchronized with the planned speech output. SMARTKOM's lip synchronization is based on a simple mapping between phonemes and visemes. A viseme is a picture of a particular mouth position of Smartakus, characterized by a specific jaw opening and lip rounding. Only plosives and diphthongs are mapped to more than one viseme.

One of the distinguishing features of SMARTKOM's modality fission is the explicit representation of generated multimodal presentations in M3L. This means that SMARTKOM ensures dialogue coherence in multimodal communication by following the design principle "no presentation without representation". The text generator provides a list of referential items that were mentioned in the last turn of the system. The display component generates an M3L representation of the current screen content, so that the discourse modeler can add the corresponding linguistic and visual objects to the discourse representation. Without such a representation of the generated multimodal presentation, anaphoric, crossmodal, and gestural references of the user could not be resolved. Thus, it is an important insight of the SMARTKOM project that a multimodal dialogue system must not only understand and represent the user's multimodal input, but also its own multimodal output.

Figure 9 shows the modality-free presentation goal that is transformed into the multimodal presentation shown in Fig. 10 by SMARTKOM's media fission component and unimodal generators and renderers. Please note that all the graphics and layout shown in Fig. 10 are generated on the fly and uniquely tailored to the dialogue

```
<presentationTask>
 <presentationGoal>
  <inform><informFocus>
   <RealizationType>list</RealizationType>
  </informFocus></inform>
  <abstractPresentationContent>
      <discourseTopic><goal>epg_browse</goal></discourseTopic>
      <informationSearch id="dim24"><tvProgram id="dim23">
          <broadcast><timeDeictic id="dim16">now</timeDeictic>
                <between>2003-03-20T19:42:32 2003-03-20T22:00:00</between>
            <channel><channel id="dim13"/></channel>
           </broadcast></tvProgram>
       </informationSearch>
    <result><event>
       <pieceOfInformation>
        <tvProgram id="ap_3">
          <broadcast><beginTime>2003-03-20T19:50:00</beginTime>
                      <endTime>2003-03-20T19:55:00</endTime>
                     <avMedium><title>Today's Stock News</title></avMedium>
                      <channel>ARD</channel>
          </broadcast>.....</event>
    </result>
</presentationGoal>
</presentationTask>
```

Fig. 9. A fragment of a presentation goal, as specified in M3L

situation, i.e., nothing is canned or preprogrammed. The presentation goal shown in Fig. 9 is coded in M3L and indicates that a list of broadcasts should be presented to the user. Since there is enough screen space available and there are no active constraints on using graphical output, the strategy operators applied by the presentation planner lead to a graphical layout of the list of broadcasts. In an eyes-busy situation (e.g., when the user is driving a car), SMARTKOM would decide that Smartakus should read the list of retrieved broadcasts to the user. This shows that SMARTKOM's modality fission process is highly context aware and produces tailored multimodal presentations.

The presentation planner decides that the channel should be rendered as an icon, and that only the starting time and the title of the individual TV item should be mentioned in the final presentation.

In the next section, we show how the visual, gestural and linguistic context stored in a multimodal discourse model can be used to resolve crossmodal anaphora. We will use the following dialogue excerpt as an example:

1. User: I would like to go to the movies tonight.
2. Smartakus: [displays a list of movie titles] This is a list of films showing in Heidelberg.

Fig. 10. A dynamically generated multimodal presentation based on a presentation goal

3. User: Hmm, none of these films seem to be interesting ... Please show me the TV program.
4. Smartakus: [displays a TV listing] Here [points to the listing] is a listing of tonight's TV broadcasts (see Fig. 10).
5. User: Please tape the third one!

9 A Three-Tiered Multimodal Discourse Model

Discourse models for spoken dialogue systems store information about previously mentioned discourse referents for reference resolution. However, in a multimodal dialogue system like SMARTKOM, reference resolution relies not only on verbalized, but also on visualized information. A multimodal discourse model must account for entities not explicitly mentioned (but understood) in a discourse, by exploiting the verbal, the visual and the conceptual context. Thus, SMARTKOM's multimodal discourse representation keeps a record of all objects visible on the screen and the spatial relationships between them.

An important task for a multimodal discourse model is the support of crossmodal reference resolution. SMARTKOM uses a three-tiered representation of multimodal discourse, consisting of a domain layer, a discourse layer, and a modality layer. The modality layer consists of linguistic, visual, and gestural objects that are linked to the corresponding discourse objects. Each discourse object can have various surface realizations on the modality layer. Finally, the domain layer links discourse objects with instances of SMARTKOM's ontology-based domain model (Löckelt et al.,

2002). SMARTKOM's three-tiered discourse representation makes it possible to resolve anaphora with nonlinguistic antecedents. SMARTKOM is able to deal with multimodal one-anaphora (e.g., "the third one") and multimodal ordinals ("the third broadcast in the list").

SMARTKOM's multimodal discourse model extends the three-tiered context representation of LuperFoy (1991) by generalizing the linguistic layer to that of a modality layer (see Fig. 11). An object at the modality layer encapsulates information about the concrete realization of a referential object depending on the modality of presentation (e.g., linguistic, gestural, visual). Another extension is that objects at the discourse layer may be complex compositions that consist of several other discourse objects (Salmon-Alt, 2001). For example, the user may refer to an itemized list shown on SMARTKOM's screen as a whole, or he may refer to specific items displayed in the list. In sum, SMARTKOM's multimodal discourse model provides a unified representation of discourse objects introduced by different modalities, as a sound basis for crossmodal reference resolution.

The modality layer of SMARTKOM's multimodal discourse model contains three types of modality objects:

- Linguistic Objects (LOs): For each occurrence of a referring expression in SMARTKOM's input or output, one LO is added.
- Visual Objects (VOs): For each visual presentation of a referrable entity, one VO is added.
- Gesture Objects (GOs): For each gesture performed either by the user or the system, a GO is added.

Each modality object is linked to a corresponding discourse object. The central layer of the discourse model is the discourse object layer. A Discourse Object (DO) represents a concept that can serve as a candidate for referring expressions, including objects, events, states and collections of objects. When a concept is newly introduced by a multimodal communicative act of the user or the system, a DO is created. For each concept introduced during a dialogue, there exists only one DO, regardless of how many modality objects mention this concept.

The compositional information for the particular DOs that represent collections of objects, is provided by partitions (Salmon-Alt, 2001). A partition provides information about possible decompositions of a domain object. Such partitions are based either on perceptual information (e.g., a set of movie titles visible on the screen) or discourse information (e.g., "Do you have more information about the first and the second movie?" in the context of a list of movie titles presented on the screen). Each element of a partition is a pointer to another DO, representing a member of the collection. The elements of a partition are distinguishable from one another by at least one differentiation criterion like their relative position on the screen, their size, or color. For instance, the TV listing shown in Fig. 10 is one DO that introduces 13 new DOs corresponding to particular broadcasts.

The domain object layer provides a mapping between a DO and instances of the domain model. The instances in the domain model are Ontological Objects (OO) that

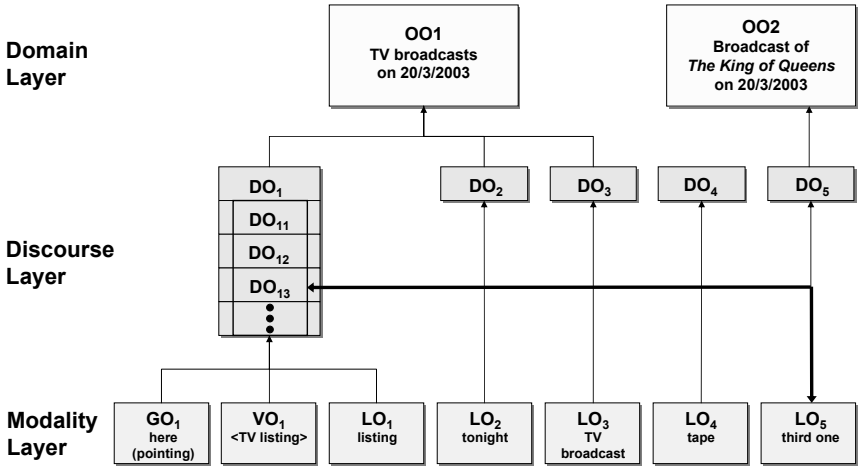

Fig. 11. An excerpt from SMARTKOM's multimodal discourse model

provide a semantic representation of actions, processes, and objects. SMARTKOM's domain model is described in the ontology language OIL (Fensel et al., 2001).

Let us discuss an example of SMARTKOM's methodology for multimodal discourse modeling. The combination of a gesture, an utterance, and a graphical display that is generated by SMARTKOM's presentation planner (see Fig. 10) creates the gestural object GO_1, the visual object VO_1 and the linguistic object LO_1 (see Fig. 11). These three objects at the modality layer are all linked to the same discourse object DO_1 that refers to the ontological object OO1 at the domain layer. Note that DO_1 is composed of 13 subobjects. One of these subobjects is DO_{13}, which refers to OO2, the broadcast of "The King of Queens" on 20 March 2003 on the ZDF channel. Although there is no linguistic antecedent for the one-anaphora "the third one", SMARTKOM can resolve the reference with the help of its multimodal discourse model. It exploits the information that the spatial layout component has rendered OO1 into a horizontal list, using the temporal order of the broadcasts as a sorting criterion. The third item in this list is DO_{13}, which refers to OO2. Thus, the cross-modal one-anaphora "the third one" is correctly resolved and linked to the broadcast of "The King of Queens" (see Fig. 11).

During the analysis of turn (3) in the dialogue excerpt above, the discourse modeler receives a set of hypotheses. These hypotheses are compared and enriched with previous discourse information, in this example stemming from (1). Although (3) has a different topic to (1) (it requests information about the cinema program, whereas (3) concerns the TV program), the temporal restriction (tonight) of the first request is propagated to the interpretation of the second request. In general, this propagation of information from one discourse state to another is obtained by comparing a current intention hypothesis with previous discourse states, and by enriching it (if possible) with consistent information. For each comparison, a score has to be computed re-

flecting how well this hypothesis fits in the current discourse state. For this purpose, the nonmonotonic OVERLAY operation (an extended probabilistic unification-like scheme, see Alexandersson and Becker (2003)) has been integrated into SMARTKOM as a central computational method for multimodal discourse processing.

10 Beyond Restricted Domains: From SmartKom to SmartWeb

Although SMARTKOM works in multiple domains (e.g., TV program guide, telecommunication assistant, travel guide), it supports only restricted-domain dialogue understanding. Our follow-up project SmartWeb (duration: 2004–2008) goes beyond SMARTKOM in supporting *open-domain question answering* using the entire Web as its knowledge base.

Recent progress in *mobile broadband communication* and *semantic Web technology* is enabling innovative mobile Internet information services, that offer much higher retrieval precision than current Web search engines like Google or Yahoo!. The goal of the SmartWeb project (Reithinger et al., 2005) is to lay the foundations for multimodal interfaces to wireless Internet terminals (e.g., smart phones, Web phones, PDAs) that offer flexible access to various Web services. The SmartWeb consortium brings together experts from various research communities: mobile Web services, intelligent user interfaces, multimodal dialogue systems, language and speech technology, information extraction, and semantic Web technologies (see http://www.smartweb-project.org).

SmartWeb is based on the fortunate confluence of three major efforts that have the potential to form the basis for the next generation of the Web. The first effort is the *Semantic Web* (Fensel et al., 2003) which provides the tools for the explicit markup of the content of webpages; the second effort is the development of *semantic Web services* which results in a Web where programs act as autonomous agents to become the producers and consumers of information and enable automation of transactions. The third important effort is information extraction from huge volumes of rich text corpora available on the Web. There has been substantial progress in extracting named entities (such as person names, dates, locations) and facts relating these entities, for example "Winner[World_Cup, Germany, 1990, Italy]", from arbitrary text.

The appeal of being able to ask a question to a mobile Internet terminal and receive an answer immediately has been renewed by the broad availability of always-on, always-available Web access, which allows users to carry the Internet in their pockets. Ideally, a multimodal dialogue system that uses the Web as its knowledge base would be able to answer a broad range of questions. Practically, the size and dynamic nature of the Web and the fact that the content of most webpages is encoded in natural language makes this an extremely difficult task. However, SmartWeb exploits the machine-understandable content of semantic webpages for intelligent question-answering as a next step beyond today's search engines. Since semantically annotated webpages are still very rare due to the time-consuming and costly manual markup, SmartWeb is using advanced language technology, information extraction

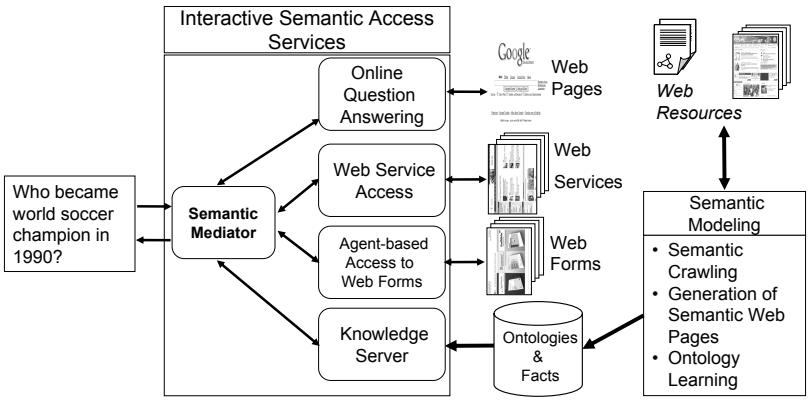

Fig. 12. The semantic mediator of SmartWeb

methods and machine learning for the automatic annotation of traditional webpages encoded in HTML or XML. SmartWeb generates such semantic webpages offline and stores the results in an ontology-based database of facts that can be accessed via a knowledge server (see Fig. 12). In addition, the semantic mediator of SmartWeb uses online question answering methods based on real-time extraction of relevant information from retrieved webpages.

But SmartWeb does not only deal with information-seeking dialogues but also with task-oriented dialogues, in which the user wants to perform a transaction via a Web service (e.g., program his navigation system to find the soccer stadium). Agent-based access to web forms allows the semantic mediator to explore the so-called Deep Web, including webbed databases, archives, dynamically created webpages and sites requiring login or registration.

SmartWeb provides a *context-aware user interface*, so that it can support the user in different roles, e.g., as a car driver, a motor biker, a pedestrian or a sports spectator. One of the demonstrators of SmartWeb is a personal guide for the 2006 FIFA World Cup in Germany, which provides mobile infotainment services to soccer fans, anywhere and anytime (see Fig. 13). The academic partners of SmartWeb are the research institutes DFKI (consortium leader), FhG FIRST, and ICSI together with university groups from Erlangen, Karlsruhe, Munich, Saarbrücken, and Stuttgart. The industrial partners of SmartWeb are BMW, DaimlerChrysler, Deutsche Telekom, and Siemens as large companies, as well as EML, Ontoprise, and Sympalog as small businesses. The German Federal Ministry of Education and Research (BMBF) is funding the SmartWeb consortium with grants totaling 13.7 million €.

11 The Roadmap for Multimodality

This book presents the foundations of multimodal dialogue systems using our fully fledged SMARTKOM system as an end-to-end working example. However, in this

Fig. 13. SmartWeb: open-domain and multimodal question-anwering

young research field many foundational questions are still open, so that intensive research on multimodality will be needed throughout the next decade. Our research roadmap for 2006–2010 shown in Fig. 14 outlines the research agenda for multimodality (Bunt et al., 2005).

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. From 2006 to 2010, in the area of models of multimodality, we envision biologically inspired intersensory coordination models, test suites and benchmarks for comparing, evaluating and validating multimodal systems, and eventually computational models of the acquisition of multimodal communication skills, among other advancements. Advanced methods will include affective, collaborative, multiparty, 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 on mobile devices. The annual international conference on multimodal interfaces (ICMI) has become the premier venue for presenting the latest research results on multimodal dialogue systems (see, e.g., Oviatt et al. (2003)).

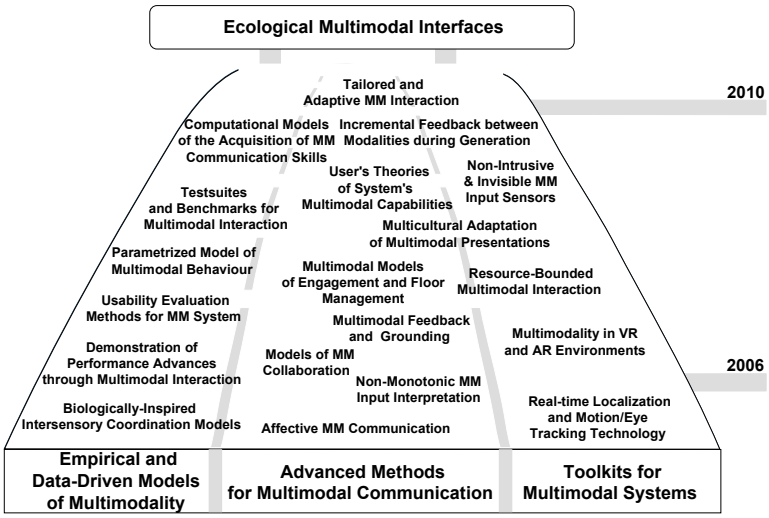

Fig. 14. Research roadmap for advanced multimodal systems

12 The Economic and Scientific Impact of SmartKom

The industrial and economic impact of the SMARTKOM project is remarkable. Up to now, 52 patents concerning SMARTKOM technologies have been filed by members of the SMARTKOM consortium, in areas such as speech recognition (13), dialogue management (10), biometrics (6), video-based interaction (3), multimodal analysis (2), and emotion recognition (2).

In the context of SMARTKOM, 59 new product releases and prototypes have surfaced during the project's life span. 29 spin-off products have been developed by the industrial partners of the SMARTKOM consortium at their own expense. The virtual mouse, which was invented by Siemens, is a typical example of such a technology transfer result. The virtual mouse has been installed in a cell phone with a camera. When the user holds a normal pen about 30 cm in front of the camera, the system recognizes the tip of the pen as a mouse pointer. A red point then appears at the tip on the display. For multimodal interaction, the user can move the pen and point to objects on the cell phone's display.

Former researchers from the SMARTKOM consortium have founded six start-up companies, including Sonicson, Eyeled, and Mineway.[1] The product spectrum of these companies includes multimodal systems for music retrieval, location-aware mobile systems, and multimodal personalization systems.

In addition to its economic impact, SMARTKOM has a broad scientific impact. The scientific results of SMARTKOM have been reported in 255 publications and 117 keynotes or invited lectures. During the project, six SMARTKOM researchers were

[1] http://www.sonicson.com, http://www.eyeled.com, and http://www.mineway.de

awarded tenured professorship. 66 young researchers have finished their master's or doctoral theses in the context of the SMARTKOM project.

SMARTKOM's MULTIPLATFORM software framework (see Sect. 6) is being used at more than 15 industrial and academic sites all over Europe and has been selected as the integration framework for the COMIC (COnversational Multimodal Interaction with Computers) project funded by the EU (Catizone et al., 2003). SMARTKOM's multimodal markup language M3L had an important impact on the definition of MMIL, which is now actively used in the ISO standardization effort towards a multimodal content representation scheme in ISO's Technical Committee 37, Subcommittee 4 "International Standards of Terminology and Language Resource Management". In addition, M3L had an obvious impact on the W3C effort towards a standard for a natural language semantics markup language (see http://www.w3.org/TR/nl-spec/).

The sharable multimodal resources collected and distributed during the SMARTKOM project will be useful beyond the project's life span, since these richly annotated corpora will be used for training, building, and evaluating components of multimodal dialogue systems in coming years. 448 multimodal Wizard-of-Oz sessions resulting in 1.6 terabytes of data have been processed and annotated (Schiel et al., 2002). The annotations contain audio transcriptions combined with gesture and emotion labeling.

Acknowledgments

The SMARTKOM project was made possible by funding from the German Federal Ministry of Education and Research (BMBF) under grant 01 IL 905. I would like to thank my SMARTKOM team at DFKI: Jan Alexandersson, Tilman Becker, Anselm Blocher (project management), Ralf Engel, Gerd Herzog (system integration), Heinz Kirchmann, Markus Löckelt, Stefan Merten, Jochen Müller, Alassane Ndiaye, Rainer Peukert, Norbert Pfleger, Peter Poller, Norbert Reithinger (module coordination), Michael Streit, Valentin Tschernomas, and our academic and industrial partners in the SMARTKOM project consortium: DaimlerChrysler AG, European Media Laboratory GmbH, Friedrich-Alexander University Erlangen-Nuremberg, International Computer Science Institute, Ludwig-Maximilians University Munich, MediaInterface GmbH, Philips GmbH, Siemens AG, Sony International (Europe) GmbH, and Stuttgart University for the excellent and very successful cooperation.

References

J. Alexandersson and T. Becker. The Formal Foundations Underlying Overlay. In: *Proc. 5th Int. Workshop on Computational Semantics (IWCS-5)*, pp. 22–36, Tilburg, The Netherlands, February 2003.

A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer. The Recognition of Emotion. In: W. Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 122–130, Berlin Heidelberg New York, 2000. Springer.

H. Bunt, M. Kipp, M. Maybury, and W. Wahlster. Fusion and Coordination for Multimodal Interactive Information Presentation. In: O. Stock and M. Zancanaro (eds.), *Multimodal Intelligent Information Presentation*, vol. 27 of *Text, Speech and Language Technology*, pp. 325–340, Berlin Heidelberg New York, 2005. Springer.

R. Catizone, A. Setzer, and Y. Wilks. Multimodal Dialogue Management in the CoMIC Project. In: *Proc. EACL-03 Workshop on "Dialogue Systems: Interaction, Adaptation and Styles of Management"*, Budapest, Hungary, April 2003. European Chapter of the Association for Computational Linguistics (EACL).

P.R. Cohen, M. Johnston, D. McGee, S.L. Oviatt, J.A. Pittman, I. Smith, L. Chen, and J. Clow. QuickSet: Multimodal Interaction for Distributed Applications. In: *Proc. 5th Int. Multimedia Conference (ACM Multimedia '97)*, pp. 31–40, Seattle, WA, 1997. ACM.

D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster (eds.). *Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential*. MIT Press, Cambridge, MA, 2003.

D. Fensel, F. van Harmelen, I. Horrocks, D.L. McGuinness, and P.F. Patel-Schneider. OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2):38–45, 2001.

I. Gurevych, S. Merten, and R. Porzel. Automatic Creation of Interface Specifications from Ontologies. In: H. Cunningham and J. Patrick (eds.), *Proc. HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*, pp. 59–66, Edmonton, Canada, 2003. Association for Computational Linguistics.

G. Herzog, H. Kirchmann, S. Merten, A. Ndiaye, P. Poller, and T. Becker. MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems. In: H. Cunningham and J. Patrick (eds.), *Proc. HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*, pp. 75–82, Edmonton, Canada, 2003. Association for Computational Linguistics.

M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. MATCH: An Architecture for Multimodal Dialogue Systems. In: *Proc. 10th ACM Int. Symposium on Advances in Geographic Information Systems*, pp. 376–383, Washington, DC, 2002.

M. Löckelt, T. Becker, N. Pfleger, and J. Alexandersson. Making Sense of Partial. In: C.M. Johan Bos, Mary Ellen Foster (ed.), *Proc. 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002)*, pp. 101–107, Edinburgh, UK, September 2002.

S. LuperFoy. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*. PhD thesis, University of Texas at Austin, December 1991.

D.L. Martin, A.J. Cheyer, and D.B. Moran. The Open Agent Architecture: A Framework for Building Distributed Software Systems. *Applied Artificial Intelligence*, 13(1–2):91–128, 1999.

M.T. Maybury and W. Wahlster. Intelligent User Interfaces: An Introduction. In: M.T. Maybury and W. Wahlster (eds.), *Readings in Intelligent User Interfaces*, pp. 1–13, San Francisco, CA, 1998. Morgan Kaufmann.

S. Oviatt. Multimodal Interfaces. In: J.A. Jacko and A. Sears (eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, pp. 286–304, Mahwah, NJ, 2003. Lawrence Erlbaum.

S.L. Oviatt, T. Darrell, M.T. Maybury, and W. Wahlster (eds.). *Proc. Int. Conf. on Multimodal Interfaces (ICMI'03)*, Vancouver, Canada, November 5–7 2003. ACM.

N. Pfleger, J. Alexandersson, and T. Becker. Scoring Functions for Overlay and Their Application in Discourse Processing. In: *Proc. KONVENS 2002*, pp. 139–146, Saarbruecken, Germany, September–October 2002.

N. Reithinger, S. Bergweiler, R. Engel, G. Herzog, N. Pfleger, M. Romanelli, and S. Sonntag. A Look Under the Hood — Design and Development of the First SmartWeb System Demonstrator. In: *Proc. Int. Conf. on Multimodal Interfaces (ICMI'05)*, pp. 159–166, Trento, Italy, 2005.

S. Salmon-Alt. Reference Resolution Within the Framework of Cognitive Grammar. In: *Int. Colloquium on Cognitive Science*, pp. 1–15, San Sebastian, Spain, May 2001.

F. Schiel, S. Steininger, and U. Türk. The SmartKom Multimodal Corpus at BAS. In: *Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, pp. 35–41, Las Palmas, Spain, 2002.

S. Seneff, R. Lau, and J. Polifroni. Organization, Communication, and Control in the Galaxy-II Conversational System. In: *Proc. EUROSPEECH-99*, pp. 1271–1274, Budapest, Hungary, 1999.

W. Wahlster. User and Discourse Models for Multimodal Communication. In: J.W. Sullivan and S.W. Tyler (eds.), *Intelligent User Interfaces*, pp. 45–67, New York, 1991. ACM.

W. Wahlster. SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions. In: *Proc. 1st Int. Workshop on Man-Machine Symbiotic Systems*, pp. 213–225, Kyoto, Japan, 2002.

W. Wahlster, E. André, W. Finkler, H.J. Profitlich, and T. Rist. Plan-Based Integration of Natural Language and Graphics Generation. *Artificial Intelligence*, 63:387–427, 1993.

W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In: *Proc. EUROSPEECH-01*, vol. 3, pp. 1547–1550, Aalborg, Denmark, September 2001.

W. Wahlster and R. Wasinger. The Anthropomorphized Product Shelf: Symmetric Multimodal Interaction with Instrumented Environments. In: E. Aarts and J. Encarnação (eds.), *True Visions: The Emergence of Ambient Intelligence*, Berlin Heidelberg New York, 2006. Springer.