Context-Sensitive Multimodal Mobile Interfaces

Speech and Gesture Based Information Seeking Interaction with Navigation Maps on Mobile Devices

Daniel Sonntag*
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
sonntag@dfki.de

ABSTRACT

Multimodal speech-based interfaces offer great opportunities when designing mobile Human Computer Interfaces (HCIs). On the other hand, even closed-domain speech applications have to deal with many different linguistic surface forms, which unfortunately results in an error-prone speech recognition step at the beginning of a potentially complex natural language understanding and processing pipeline within dialogue systems. The challenge we address is to motivate the user to use speech when asking for specific information in map-based navigation scenarios, while at the same time guiding the user to restrict to a specific vocabulary. Multimodal navigation map interface design plays the major role, where, for example, click events on a map can be accompanied by natural language questions.

Keywords

Dialogue Systems, Interaction Design, Graphical HCI, Multimodality

1. INTRODUCTION

Combining mobile navigation and information-seeking scenarios, two major questions might be considered. Firstly, how to help users express their request in the user context, and secondly, how to present the information in this context, which means the situational context and the linguistic context of an ongoing dialogue. In any case, multimedia presentations on mobile device displays play a significant role. The same applies to multimodal interface design to navigation maps, where context sensitive multimodal interpretation and context sensitive multimodal output generation should help to build natural and task-optimised mobile HCIs. The user should be able to pose questions in context using multiple modalities, and using speech as primary interaction mode. The good news to us is that with (only) a good domain-specific graphical interface based interaction design we can alleviate the affects of non-existing speech recognition and natural language understanding capabilities on mobile devices. The remainder of this document is concerned with showing examples in favour of this hypothesis in specific mobile navigation usage scenarios. We will show how we designed a map application that uses speech and gesture as input, where speech remains the dominant input modality, although gestures can be used where most appropriate. The positive side effect is that the user utterances are quite predictable in the specific map presentation context, so that simpler language model can be used for speech recognition and natural language understanding purposes.

The text is organised as follows: section 2 introduces mobile map information access on mobile devices. Section 3 is concerned with presenting our implementation of a graphical multimodal map navigation application—as an interface to the Semantic Web. We finally draw some conclusions in section 4.

2. MOBILE INFORMATION ACCESS

The left screen in figure 1 illustrates that pointing gestures as user interaction with mobile interfaces resemble hyperlink mouse clicks of traditional desktop browsers. The user quickly feels familiar with this kind of navigation and data exploration pattern. A click on a generated hyperlink is interpreted as a command to open the link, and as selection of the underlying hyperlink document to bring into visual focus. Pointing gestures can be used to complement language input as in *"How far is this town from here?"* + [Pointing gesture on Munich] (figure 1(right)). In addition, graphical user interface (GUI) interaction can be leveraged by meta dialogue which the user utters as speech commands, such as accept, reject, going back to previous maps or graphical displays. Apart from these opportunities, a major challenge is how to pose questions in context using multiple modalities. Or how to exploit the synergies between the available input and output modalities and output media types, such as navigation maps with points of interest (POIs), textual descriptions, and speech output.

*The research presented here is sponsored by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A (SmartWeb).

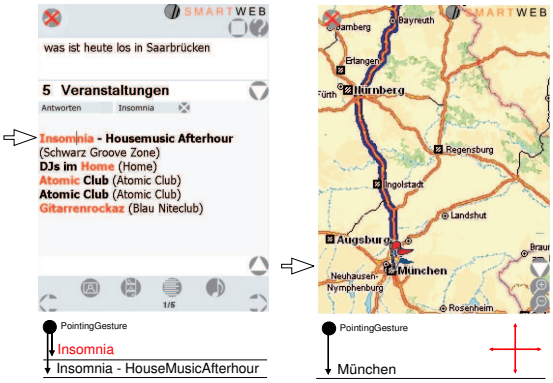

Figure 1: Pointing Gestures allow the selection of hyperlinks links, text entities, and POIs. Every pointing gesture should refer to a visual object that is transmitted to an input fusion module.

Let's imagine an application example to help answer these questions. Given that the user was correctly guided by a traffic navigation application that runs on her smartphone she takes along in the car, she arrives at her destination city, Munich. The following interaction sequence would be a quite natural and optimised dialogue in a local navigation scenario where reference resolution, map presentation, and multimodal interaction play the main roles.

U: *"Where can I find Italian Restaurants?"*

S: Shows a map with POIs and the restaurant names + synthesis: *"Restaurants are displayed"*

U: *"... and where's an ATM?"*

S: Shows a map with POIs and ATM locations nearby + synthesis: *"ATMs are displayed"*

U: Pointing gesture on a suitable ATM POI[1] + synthesis: *"How can I get there from here?"*

S: Zooms into the map and shows the route + synthesis: *"To Schiller Strasse (350 m)"*

We implemented these dialogue processing and navigation capabilities, and encountered further restrictions to be considered in the mobile user context, apart from the difficulties in understanding the speech dialogues. If you display maps with POIs in walking distance on a mobile device you will recognise that maps are too small to identify street names and POIs you are interested in. A medium resolution of 740 × 740 pixels for a map can hardly be displayed on a smartphone GUI, even in fullscreen mode. In addition, POIs are just too tiny to click when pointing gestures are legal and natural interaction means.

In order to overcome these deficiencies, we design and implement a multimodal mobile HCI solution for local navigation. The first piece of technology we need is a reliable (distributed) dialogue system for mobile environments which we can re-use for more navigation driven scenarios. We used the SmartWeb multimodal dialogue system for that purpose which we briefly introduce in the next two subsections.

2.1 Distributed Dialogue System Architecture

In distributed dialogue systems a flexible dialogue system platform is required in order to allow for true multi-session operations with multiple concurrent users of the server-side system as well as to support audio transfer and other data connections between the mobile device and a remote dialogue server. These types of systems have been developed, like the Galaxy Communicator [3] (cf. also [13, 4, 2]), In the context of the SmartWeb project[2] we developed an ontology-based dialogue system for that purpose.

The partners of the SmartWeb project share experience from earlier dialogue system projects [17, 18, 12]. We followed guidelines for multimodal interaction, as explained in [7] for example, in the development process of our first demonstrator system [5] which contains the following assets: *multimodality*, more modalities allow for more natural communication, *encapsulation*, we encapsulate the multimodal dialogue interface proper from the application, *standards*, adopting to standards opens the door to scalability, since we can re-use ours as well as other's resources, and *representation*. We base all message transfer and knowledge structures on ontology instances as an integration test of ontologies in knowledge-based natural language processing applications.

The dialogue server system platform instantiates one dialogue server for each call and connects the multimodal recogniser for speech and gesture recognition. The dialogue system instantiates and sends the requests to the Semantic Web access, which provides the umbrella for all different access methods to the Semantic Web we use. The most important processing modules connected to the dialogue system are: a speech interpretation component (SPIN), a modality fusion and discourse component (FADE), a system reaction and presentation component (REAPR), and a natural language generation module (NIPSGEN). See [15] for further information and references. The distributed dialogue system architecture for Semantic Web Service access is shown in figure 2.

We used standardised interface descriptions (EMMA[3], SSML[4], RDF[5], OWL-S[6], WSDL[7], SOAP[8], MPEG7[9]). Our representation of discourse, domain, and linguistic concepts is

[1] In Germany, some house banks do not waive fees to draw money.

[2] http://www.smartweb-project.de/start_en.html
[3] http://www.w3.org/TR/emma
[4] http://www.w3.org/TR/speech-synthesis
[5] http://www.w3.org/TR/rdf-primer
[6] http://www.w3.org/Submission/OWL-S
[7] http://www.w3.org/TR/wsdl
[8] http://www.w3.org/TR/soap
[9] http://www.chiariglione.org/mpeg

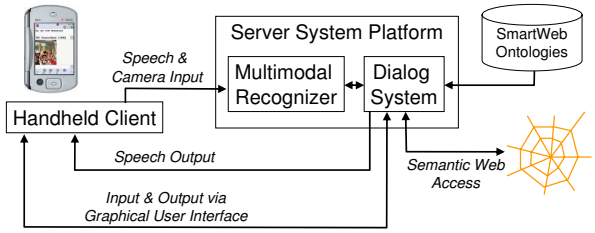

Figure 2: SmartWeb's distributed mobile dialogue system architecture for accessing Semantic Web Services. Recorded speech is transferred to a server platform where a multimodal recogniser, a dialogue system, and the connection to the ontologies and the Semantic Web Services is realised. Camera input can also be interpreted in more question answering based application scenarios.

the SWIntO (**S**mart**W**eb **Int**egrated **O**ntology)[6].

SWIntO integrates question answering specific knowledge in a discourse ontology and integrates multimedia information in a media ontology. The data exchange between the dialogue system and the backend server is RDF-based. We believe that consistent and homogeneous presentation of multimedia answers such as navigation maps is the result of ontological representation of common sense interaction and presentation knowledge, e.g., interaction patterns in interactive mobile domains (cf. [14]).

Unlike context-sensitive information retrieval, and database retrieval, we address retrieval from Semantic Web Services, which constitute our information repositories. Of course, natural language processing for deeper understanding of user utterances is an important step in our process pipeline. Another important step is the connection to the Web Services. To connect to these services we developed a semantic representation formalism based on OWL-S and a service composition component able to interpret an ontological user query connected to the T-Info (DTAG) Web Services[10] offered by Deutsche Telekom AG (see also [1]). A Web Service composer addresses over twenty different (commercial) external services from navigation to event information and product information (books, movies). In the context of information seeking interaction with navigation maps, we here extend the user interaction possibilities by results obtained from Semantic Web Services that deliver image maps with POIs, i.e., hotel-, gas station-, and bank location services.

2.2 Discourse and Situational Context

Challenges for the evaluation of emerging Human Computing applications [10] traces back to challenges in multimodal dialogue processing, such as error-prone perception and integration of multimodal input channels [8, 19]. All recognised user actions should be processed with respect to their situational and discourse context. A user is thus not required to pose separate and unconnected questions. The task of the fusion module *FADE* is to integrate the verbal and non-verbal user contributions into a coherent multimodal representation to be enriched by contextual information, e.g., resolution of referring and elliptical expressions [9].

The situational context is maintained by another component called *SitCom*[11], which provides GPS co-ordinates information and other context information such as weather conditions, and specific user preferences. In our example of the application scenario, the *SitCom* module provides the respective co-ordinates for the user reference *"from here"*. Dialogue-based relevance-feedback queries are often deictic in nature, consider for example: *"Show me more like this"*. Fusion of multimodal input, where the user clicks on items first or utters co-referenced or deictic items, are special topics of interest. An example semantic query with successful *"from here"* reference resolution (deictic, linguistic, and geographic positioning resolution) of the user enquiry: *"Where can I find ATMs not far from here?"* is the following semantic typed feature structure, which we send as query to the Semantic Web Service subsystem.

```
[ Query
  text: Where can I find ATMs not far from here?
  dialogueAct: [discourse#Question]
  focus:
    [ Focus
        focusMediumType:  [ mpeg7#Text]
        focusMediumType:  [ mpeg7#Image]
        varContext:
           [
             contextObject: #1
           ]
        varName:X
      ]
   content:
       [ QEPattern
              patternArg:
              #1 [ [sumo#POI:
                navigation#Cashpoint
                     ]
                     . . .
                      [sumo#Map]


         inCity: [Berlin]
                      [sumo#centerAddress:
                sumo#GEOPOSITION:
                           [N52f31.19' E13f24.69' (WGS84)]
        ]


                [context#vehicleState:[Car] . . .]


 ]
```

In the following we focus on the navigation map front end user interface we developed for context-sensitive multimodal input fusion capabilities. We first explain the multimodal presentation planning and design guidelines.

2.3 Presentation Design Guidelines

Presentation planning manages the dialogical interaction for the supported dialogue phenomena such as flexible turn-taking, incremental result presentation, and multimodal fusion of system output. One of the challenges is to allow for selection by voice in deictic utterances like *"Show me this and that "* and *"Where can I find X nearby? / not far from here?"* as introduced in the dialogue example. Multimodal

[10] http://services.t-info.de/soap.index.jsp

feedback to multimodal user dialogue acts are the core requirements, i.e.,

- to produce useful reactions and to give hints to the user or examples that the use of supported terminology is not insisted, but at least directed.
- to keep acoustic messages short and simple. Speech synthesis belongs to the more obtrusive output channels. In the context of handheld devices, this guideline should be followed with even more strictness, since the capabilities of the handheld device loudspeaker are very limited, and the social and situation context in which the handheld is used (confer modality busy setting) does not permit longer speech syntheses in many cases.[11]
- to correspond the synthesis to a smaller fragment of text which is displayed concurrently.
- to deal with layout as a rhetorical force, influencing the intentional and attentional state of the discourse participants [16].

For the graphics and GUI surface we decided on Flash-movies to be locally played on the PDA device as user interface. Flash MX 2004 / Flash8 is one of the industry standard tools for creating effective rich content and applications.

3. GRAPHICAL MAP INTERACTION

In this section we present our multimodal interaction design and its implementation on a smartphone. Suppose the user poses a questions in context using speech: *"Where can I find an ATM nearby?"*. Figure 3 shows the user interface and map interaction fragment in a real-estate user scenario with commercial navigation Web Services involved in response to this question.

The interactions shown in figure 3 represent the starting point for our speech-based multimodal interaction design. The goal of multimodal interaction design as presented here is to allow for interactions as a combination of voice and graphical user interface in a generic setting. Generic setting means that the input modality mix can be reduced in a modality busy setting, e.g., when the phone is on the ear, or when it is too loud around the user. In general, modality busy means that a modality is not available, not commutable, or not appropriate in a specific situation. In any case, much attention is drawn to the speech input (first and foremost for posing the natural language query or command), since the user interaction device is a phone and speech input most natural.

Concerning multimodal map result presentations, specific circumstances are expected in map interaction scenarios: Many results are not naturally presented auditory, which means that we try to exploit the graphical display as much as possible. Accordingly, all results will be presented on

[11] Although people are using more and more headphones, which shifts the situational context, we hold that longer speech syntheses (in lower quality) are distracting in navigation scenarios.

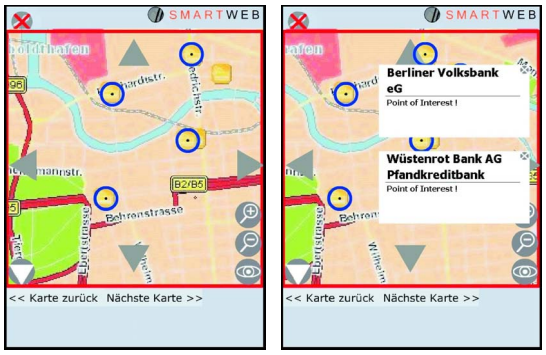

Figure 4: POI and additional textual POI info selection

screen at least. This also means that every multimodal dialogue act, such as presenting answers or to give immediate feedback on a user command or query, is realised as graphical presentation. On this note, we propose an integration of graphics and speech output:

Having received a navigation map with POIs, we address the question, how the map, or parts of it, fits best in the user context, and how additional (text-based) information about the POIs should be presented, either textually or auditory. The last point concerns both the aspects of map area selection (first occurrence in figure 3(area selection)), content selection, and content filtering.

As can be seen in figure 4, POIs can be clicked in order to query for additional server-side information. As query feedback, we draw a blue circle if the POI can be referenced to a Web Service map annotation. In addition, we synthesise the POI names which are *"Berliner Volksbank"* and *"Wuestenrotbank"*, respectively. The user is then able to click on the POI again in order to display additional textual information about the POIs. The *Point of Interest!* subtext can be filled according to any additional information obtained from the Web Services. An example for additional POI information display will be provided in figure 6 further down.

Composite multimodal interaction does not only evoke new challenges on multimodal interaction design, but also new possibilities on the interpretation of user input—allowing for mutual disambiguation of modalities. For example, if the discourse context is not clear enough to dereference the starting and end point of the phrase *"How to get to Wilhelm Street"*, or the automatic speech recognition (ASR) component fails to recognise *"Wilhelm Street"* as a valid out-of-vocabulary word, the pointing co-ordinates on the touch screen can be used to calculate the desired start and end point.

The strength of multimodal discourse acts and output generation in the form of multimedia presentations can be demonstrated by the following example in figure 5. (**U1**) takes linguistic and non-linguistic user context into account to interpret the question and the word *"nearby"* in particular. The system reacts with (**S1**), thereby interpreting the vicin-

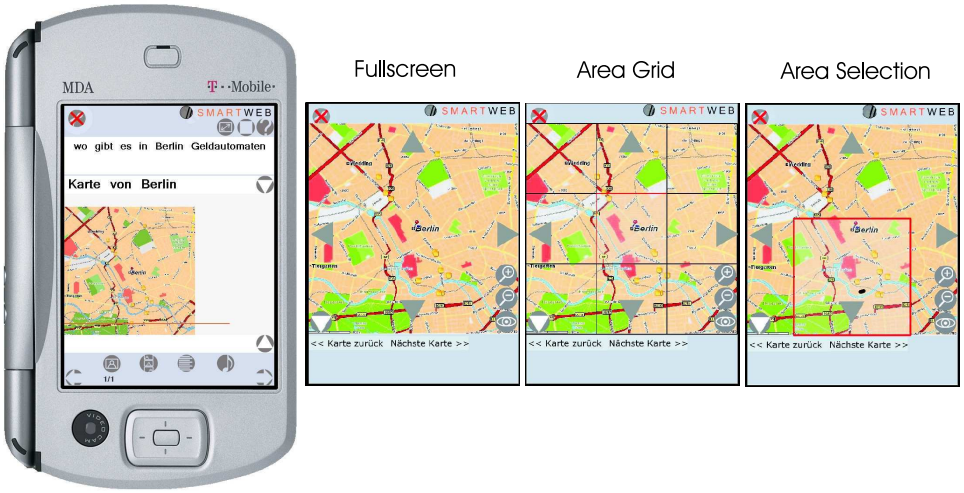

Figure 3: (Left) Graphical user interface and the response to the question *"Where can I find an ATM nearby (in Berlin)?" : "Map of Berlin"*, **which shows an area overview according to the user GPS co-ordinates. (Fullscreen) The user can activate a special fullscreen mode. (Area Grid) In order to zoom into specific areas, the user clicks on the touchscreen. (Area Selection) animates the selection of a specific area to be displayed in a specific resolution and size that motivates further click events on the screen.**

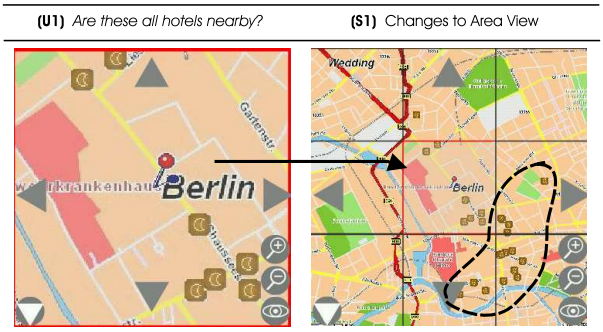

Figure 5: Multimedia presentation and transition as crucial part of a multimodal response dialogue act. The dashed line illustrates the set of additional hotels visible in Area View.

ity on an ontological basis. In addition, the larger POI map is searched for additional hotel POI occurrences, and with the help of a simple heuristic, the new area view can be successfully computed and activated. The new area activation is thereby interpreted by the user as the multimedia response to her enquiry. In combination with a result synthesis *"Further hotels found nearby"* this kind of interaction is both multimodal and effective.

In the context of the Web Service access we additionally realised the following interaction mode depicted in figure 6. This interaction mode represents the most complex one we realised. The user is engaged in a complex POI and map interaction task, whereby a more traditional navigation task is combined with user initiative information-seeking and system-initiative multimedia map presentation behaviour.

In (**U2**) the user combines a more complex drawing action with a speech based, drawing- and user context-dependent question. In (**S2**) the dialogue systems gives immediate feedback on the drawing action. In addition, in case that the connected Web Services deliver the desired (comparative) information, it is synthesised as feedback on the speech inquiry realising a symmetric multimodal interaction mode.

When the user clicks on the touchscreen, we respond by a graphical transition at least. When the user speaks, we at least respond with a short auditory message. If the user interacts with both speech and pointing gesture, we use both graphical accentuations and speech syntheses. We follow this principle for the navigation task: In (**U3**) the user clicks on a POI and gets additional text information in a graphical box. In (**U4**) the user chooses a POI and asks for route information by a context-dependent short question. The route is then calculated according to the context information provided by the *SitCom* module (GPS location, pedestrian or car driver as in the query example in section 2.2). The system then responds by (**S3**), the calculated route which is synthesised and (**S4**), the graphical presentation of the corresponding route on the navigation map.[12]

The reason why the multimodal results are presented sequentially is a technical one. The textual route descriptions can be obtained from the Web Services within ten seconds before the calculated map area can be downloaded. This takes another five to ten seconds time. We observed that the incremental presentation of this kind of navigation re-

[12] A textual route description can also be obtained from the Semantic Web Services. It can be displayed in a text field that occurs instead of the media screen where the map is displayed in the example of the interaction device and GUI above.

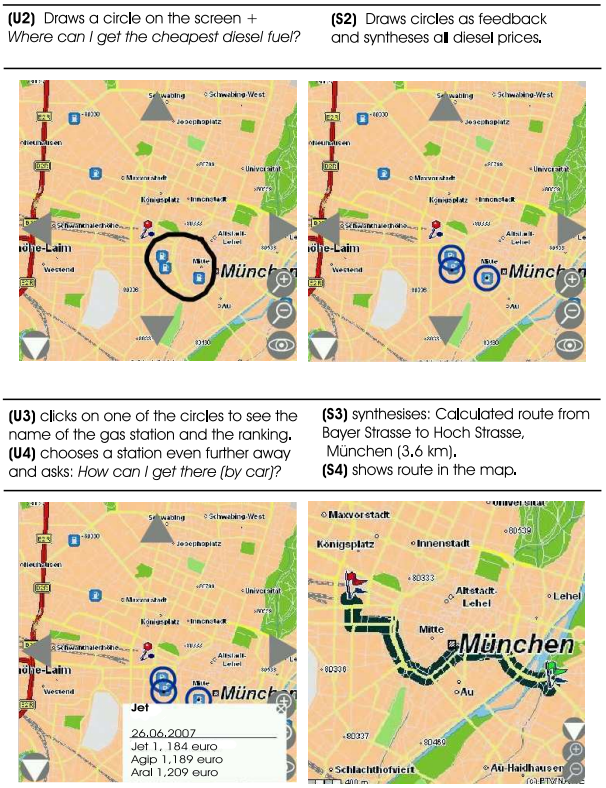

Figure 6: Speech-based navigation map interaction, information seeking, and multimedia presentation are combined into a symmetric multimodal interaction mode. User utterances are quite predictable in the GUI presentation and interaction context.

sult is perceived as natural, not at least because the synthesis bridges the time gap between the different retrieval times. In this way we avoid additional temporal synchronisation requirements for different output modalities.

4. CONCLUSION

We presented speech and gesture based interaction with navigation maps on mobile devices as an instance of a context-sensitive multimodal mobile interfaces which can be used in a context-sensitive information-seeking scenario. We described how to help users express their requests by multimodal interaction, and how to present the information in the user context, which means to consider both the situational and the linguistic context of an ongoing dialogue. The multimodal information is then presented according to the user initiative summarised in the symmetric multimodal presentation behaviour of table 1. The welcome side effect is that the user utterances are quite predictable in the specific map presentation context, so that a simple language model can be used for speech recognition and natural language understanding purposes, whereas the application in shifting contexts or similar domains is conceivable. This way we hope to contribute in developing practical solutions for real-world applications.

Table 1: User initiative symmetric multimodal system reaction and presentation behaviour

User	System
Pointing gesture	Graphical display
Speech input	Result synthesis
Speech and gesture	Speech followed by graphics
Gesture and speech	Speech and concurrent graphics

Extensions to our multimodal mobile HCI as presented here are editing functions via concurrent pen and voice, to extend symmetric multimodal result presentations to symmetric multimodal query correction. More robustness and flexibility in speech recognition and understanding is much appreciated just the same. Our future investigations will explore more fine-grained co-ordination and synchronisation of multimodal presentations in mobile environments.

ACKNOWLEDGEMENTS

The research presented here is sponsored by the German Ministry of Research and Technology (BMBF) under grant 01IMD01A (SmartWeb). We thank our student assistants for implementation support, Matthieu Deru in particular, and the project partners. The reviewers and Norbert Reithinger provided useful comments on an earlier version. The responsibility for this paper lies with the author.

5. REFERENCES

[1] A. Ankolekar, P. Hitzler, H. Lewen, D. Oberle, and R. Studer. Integrating semantic web services for mobile access. In *Proceedings of 3rd European Semantic Web Conference (ESWC 2006)*, 2006.

[2] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10, 2004. Special issue on Software Architecture for Language Engineering.

[3] A. J. Cheyer and D. L. Martin. The Open Agent Architecture. *Autonomous Agents and Multi-Agent Systems*, 4(1–2):143–148, 2001.

[4] G. Herzog, A. Ndiaye, S. Merten, H. Kirchmann, T. Becker, and P. Poller. Large-scale Software Integration for Spoken Language and Multimodal Dialog Systems. *Natural Language Engineering*, 10, 2004. Special issue on Software Architecture for Language Engineering.

[5] Norbert Reithinger and Simon Bergweiler and Ralf Engel and Gerd Herzog and Norbert Pfleger and Massimo Romanelli and Daniel Sonntag. A LooK Under the Hood. Design and Development of the First SmartWeb Demonstrator. In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI)*, Trento, Italy, 2005.

[6] D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, M. Sintek, M. Kiesel, B. Mougouie, S. Vembu, S. Baumann, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, B. Loos, R. Porzel, H.-P. Zorn, V. Micelli, C. Schmidt, M. Weiten, F. Burkhardt, and J. Zhou. DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology). Technical report,

AIFB, Karlsruhe, July 2006.

[7] S. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[8] S. Oviatt. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chapter Multimodal Interfaces, pages 286–304. Lawrence Erlbaum Assoc., 2003.

[9] N. Pfleger. Fade - an integrated approach to multimodal fusion and discourse processing. In *Proceedings of the Dotoral Spotlight at ICMI 2005*, Trento, Italy, 2005.

[10] R. Poppe and R. Rienks. Evaluating the future of hci: Challenges for the evaluation of upcoming applications. In *Proceedings of the International Workshop on Artificial Intelligence for Human Computing at the International Joint Conference on Artificial Intelligence IJCAI'07*, pages 89–96, Hyderabad, India, 2007.

[11] R. Porzel, H.-P. Zorn, B. Loos, and R. Malaka. Towards a Separation of Pragmatic Knowledge and Contextual Information. In *Proceedings of ECAI 06 Workshop on Contexts and Ontologies*, Riva del Garda, Italy, 2006.

[12] N. Reithinger, D. Fedeler, A. Kumar, C. Lauer, E. Pecourt, and L. Romary. MIAMM - A Multimodal Dialogue System Using Haptics. In J. van Kuppevelt, L. Dybkjaer, and N. O. Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*. Springer, 2005.

[13] S. Seneff, R. Lau, and J. Polifroni. Organization, Communication, and Control in the Galaxy-II Conversational System. In *Proc. of Eurospeech'99*, pages 1271–1274, Budapest, Hungary, 1999.

[14] D. Sonntag. Towards interaction ontologies for mobile devices accessing the semantic web - pattern languages for open domain information providing multimodal dialogue systems. In *Proceedings of the workshop on Artificial Intelligence in Mobile Systems (AIMS). 2005 at MobileHCI*, Salzburg, 2005.

[15] D. Sonntag, R. Engel, G. Herzog, A. Pfalzgraf, N. Pfleger, M. Romanelli, and N. Reithinger. *Artificial Intelligence for Human Computing*, volume 4451 of *LNAI*, chapter SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. Springer, Berlin, 2007.

[16] W. Wahlster. Planning multimodal discourse. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 95–96, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[17] W. Wahlster, editor. *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Springer, 2000.

[18] W. Wahlster. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In R. Krahl and D. Günther, editors, *Proc. of the Human Computer Interaction Status Conference 2003*, pages 47–62, Berlin, Germany, 2003. DLR.

[19] W. Wahlster. Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In *KI*, pages 1–18, 2003.