GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

SmartWeb Handheld Interaction (V 1.1): General Interactions and Result Display for User-System Multimodal Dialogue

Question answering (QA) has become one of the fastest growing topics in computational linguistics and information access. In this context, SmartWeb (http://www.smartweb-projekt.de/) was a large-scale German research project that aimed to provide intuitive multimodal access to a rich selection of Web-based information services. In one of the main scenarios, the user interacts with a smartphone client interface, and asks natural language questions to access the Semantic Web. The demonstrator systems were developed between 2004 and 2007 by partners from academia and industry. This document provides the interaction storyboard (the multimodal design of SmartWeb handheld's interaction), and a description of the actual implementation.

We decided to publish this technical document in a second version in the context of the THESEUS project (http://www.theseus-programm.de) since this SmartWeb document provided many suggestions for the THESEUS usability guidelines (http://www.dfki.de/~sonntag/interactiondesign.htm) and the implementation of the THESEUS TEXO demonstrators. Theseus is the German flagship project on the Internet of Services, where the user can delegate complex tasks to dynamically composed semantic web services by utilizing multimodal interaction combining speech and multi-touch input on advanced smartphones.

More information on SmartWeb's technical question-answering software architecture and the underlying multimodal dialogue system, which is further developed and commercialized in the THESEUS project, can be found in the book: *Ontologies and Adaptivity in Dialogue for Question Answering*.

Daniel Sonntag
Norbert Reithinger

DFKI GmbH

Technisches Dokument Nr. 5

December 2009

V 1.0 February 2007
V 1.1 December 2009

Daniel Sonntag
Norbert Reithinger

DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken

Tel.: (0681) 302-5254
Tel.: (0681) 302-5346
Fax: (0681) 302-5020

E-Mail: sonntag@dfki.de
bert@dfki.de

Dieses Technische Dokument gehört zu Teilprojekt 2: Verteilte Multimodale Dialogkomponente für das mobile Endgerät

Das diesem Technischen Dokument zugrunde liegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IMD 01 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

1 Introduction

SMARTWEB lays the foundations for multimodal user interfaces for distributed and composable *Semantic Web* access on mobile devices. http://smartweb.dfki.de/ provides further information. This document describes the general interaction possibilities for the SmartWeb Handheld client, in particular the result display and the user-system multimodal dialogue. We concentrate particularly on the interaction possibilities of **SmartWeb System Integration 1.0**.

The interaction design and development process was supported by iterative prototyping, close involvement of potential users, and informal user testing.

Since this is an internal and not a public document, we refrain from citing sources in the text. Instead, please refer to the references at the end of this report. This document describes past work within the SmartWeb project and is due to revision without notice. It should be read jointly with Technical Document 2 about the Interaction Storyboard, and Technical Document 3 about the architecture of the dialogue components.

The latest revision of the following dialogue components can be found at the dedicated wiki pages within the SmartWeb Developer Portal at http://smartweb.dfki.de/sys/trac/wiki:

- Information Hub for Server-side Dialogue Processing,
- Graphical User Interface Application for Handheld Dialogue Client,
- Reaction and Presentation Component for Server-side Dialogue Processing,
- Discourse Ontology,
- Server-side Multimodal Dialogue Manager,
- Server-side Media Fusion and Discourse Processing,
- Server-side Language Understanding,
- Adapted Knowledge Sources for Server-side Speech Recognition,
- Server-side Text Generation, and
- the Integrated Demonstrator.

The current version of the self-running SmartWeb presentation can be downloaded from http://smartweb.dfki.de/Intro_Demo/start.html.

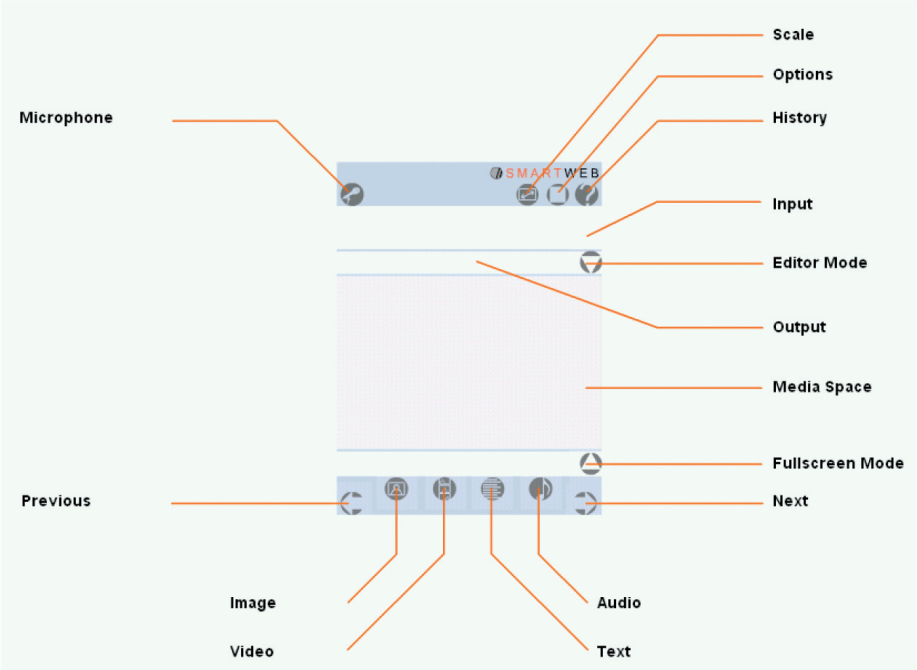
1.1 General Introduction to Interactions

The personal device for the SmartWeb Handheld scenario is T-Mobile's PDA (personal digital assistant), the MDA 4 / MDA pro. This device offers several possibilities for interaction: the microphone, the stylus (for both mouse replacement and handwriting recognition), the *Qwertz* keypad, as well as several programmable buttons, mainly on the right side of the device. The MDA 3, on which the first storyboard was based, differs mainly in terms of the display resolution and the programmable button on the front side.

The goal of this project is to allow for interactions in the form of 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 at the ear, or when the user's location is too loud. In general, "modality busy" means that any given modality is not available, not commutable, or not appropriate in a specific situation. In the scenario we model, the interaction is mainly provided by voice input and output, and interaction requires a more powerful client device with embedded speech recognition capabilities. In any case, much attention will be drawn to speech input, since the user interaction device is a phone and speech input is, therefore, most natural. For result presentation, different circumstances are expected. Many results are not naturally presented in an auditory manner, which means that we plan to exploit the graphical display as much as possible. Accordingly, all results will be presented on screen at least.

If multiple modes are used for input, multimodal discourse phenomena such as multimodal deixis, cross-modal reference, and multimodal turn-taking must be considered and modelled. The inter-

action is to be characterised as composite multimodal and the challenge is to fuse modalities to composite input in order to represent the user's intention. The composite multimodal interaction does not only evoke new challenges on interaction design, but also provides new possibilities for the interpretation of user input – allowing for mutual disambiguation of modalities. Composite input is input received on multiple modalities at the same time and treated as a single, integrated compound input by downstream processes (see, e.g., W3C Multimodal Interaction Requirements, http://www.w3.org/TR/mmi-reqs/).



SmartWeb's main Graphical User Interface (GUI) is shown in the picture to the upper right. We implemented the rendering surface according to the storyboard requirements explained in Technical Document 2; the rendering surface is roughly segmented into three sections, for input, output, and control, respectively. The input/editor field can be used to write questions. In case of active speech recognition, the best speech interpretation is shown in this input/editor field for visual inspection (user perceptual feedback), and editing possibilities. Section 2 comprises of the output field and the media space for images, videos, and longer text documents. This input-output field should reflect the SmartWeb general Question Answering (QA) paradigm in the visual surface.

QA can be defined as the task of finding answers to natural language questions by searching large document collections. Unlike information retrieval systems, question answering systems do not retrieve documents, but provide short, relevant answers located in small fragments of text instead. Accordingly, the input field is for posing a complex question, and the output field should display a short relevant answer to it. This answer can also be seen as a headline to more complex (multimedia) answers, when a greater amount of information should be displayed than a few words, or multimedia content should be presented. For example, the question "Show me a picture of the 2006 World Cup's mascots" results in displaying the name *Goleo* in the answer field, followed by the image, both of which are relevant to the multimodal answer (see graphic above).

Speech synthesis also plays a major role in presenting results. We decided to use speech synthesis to both notify the user that SmartWeb obtained an answer and to present content information about the answer concurrently. The output we generated for the output field seems to provide the most suitable information for synthesis. In some cases, the answer field information is a bit shorter than the synthesis. We display 'traffic condition', for example, but synthesise 'traffic condition between Aachen and Bonn'.

2 Interaction, Communication, and Synchronisation

Interaction in a multimodal system ties aspects of communication and synchronisation closely together. All interaction channels, most prominently the output channels, must be synchronised in order to render the multimodal information to be presented. Here we address this topic for the system results in particular.

2.1 Requirements

We will now describe some important communication and synchronisation considerations in the context of interactions. These take place in the dialogue until the result is presented by the system. Keep in mind that many communication aspects are dependent on presentational aspects:

- **Display unrecognised tokens in a different colour**. In case of low confidence values in recognition, i.e., out-of-vocabulary (OOV) tokens are visible in red text colour, the selection of an OOV word activates a drop-down list of all OOV candidates for selection. The first candidate is the displayed red word token.

- **Audio repetition of paraphrased query** is available, possibility of interruption (editing) the audio output as well. Editing should be allowed via speech, pen or keyboard (The paraphrased query is displayed on screen anyhow.).

- **Pull-down menu for word hypotheses in correction mode**. A precondition is that a confusion set can either be produced by lexical knowledge of the dialogue system, or by a built-in functionality of the speech recognition system. This assumes that the word hypothesis graph can be backed off to unigram probabilities (single word probabilities). Audio correction mode is also interesting in this regard, for example, if the speech recognition system is adjustable to run for single words to be uttered. In SmartWeb we do not use dynamic ASR grammars; instead, short meta commandos such as 'further', 'left', or 'as table view' are included into one general speech recognition grammar. For word hypothesis we pursue a graphical correction selection approach instead of audio correction.

- **Incremental display**: (1) for later incoming results, (2) for relevance feedback queries.

- **Higher level selection modus**, primarily for result lists: (1) quick pen list selection, (2) confirmations/approval (OK) also by onscreen touch on dynamically presented items such as images and table cells.

- **Display of result confidences**: Either low level hit/source size information or higher level confidence display at best with symbols. In any case, background statistics should be made available, e.g., as a link. For factoid questions in the open question answering scenario, best hits count most. This leads to the question of which (important) background information is to be provided for the user for result explanation and provenance.

- **User can approve selection/results by voice**. A challenging topic will be to allow for selection by voice in deictic utterances. "Show me this and that …" Dialogue-based relevance-feedback queries are often deictic in nature, such as: "Show me more like this." Fusion of multimodal input, where the user clicks on items first, is a related topic of interest.

- **Different displays for different result types**: reactions, intermediate results, final results, status information, background information. In addition (and orthogonally), the display depends on the information content itself to be rendered:the type of the question, the media of available resources, the cardinality of

results, the size of single result items (consider the different forms of semantic/syntactic web results, for example). We employ different graphical sources for different presentation requirements of the different result sets. Most of the interaction and presentation metaphors are represented by the interaction pattern branch of the discourse ontology.

• **Synchronisation of speech synthesis and textual GUI results**: Speech synthesis belongs to the more obtrusive output channels. General guidelines recommend keeping acoustic messages short and simple. In the context of handheld devices, this guideline should be followed even more strictly, since the handheld device loudspeaker has very limited capabilities, and the social and situational context during its use (confer modality busy setting) often does not permit longer speech syntheses. Hence, the synthesis should correspond to a smaller fragment of text which is displayed concurrently, and which can be seen on screen even after synthesis. In order to synchronise speech synthesis and GUI result presentation, we implement a notification framework that starts the synthesis when the output/answer field has been rendered. In this way the speech synthesis component can be synchronised with the dialogue system-internal processing. As it turns out, suitable timing parameters for the start of the synthesis depend heavily on the server machine the synthesis is processed on, and where the GUI is being started. We optimised the synchronisation behaviour for the demonstrator system using DFKI's demonstration servers and the MDA4 handheld device.

2.2 Implementations

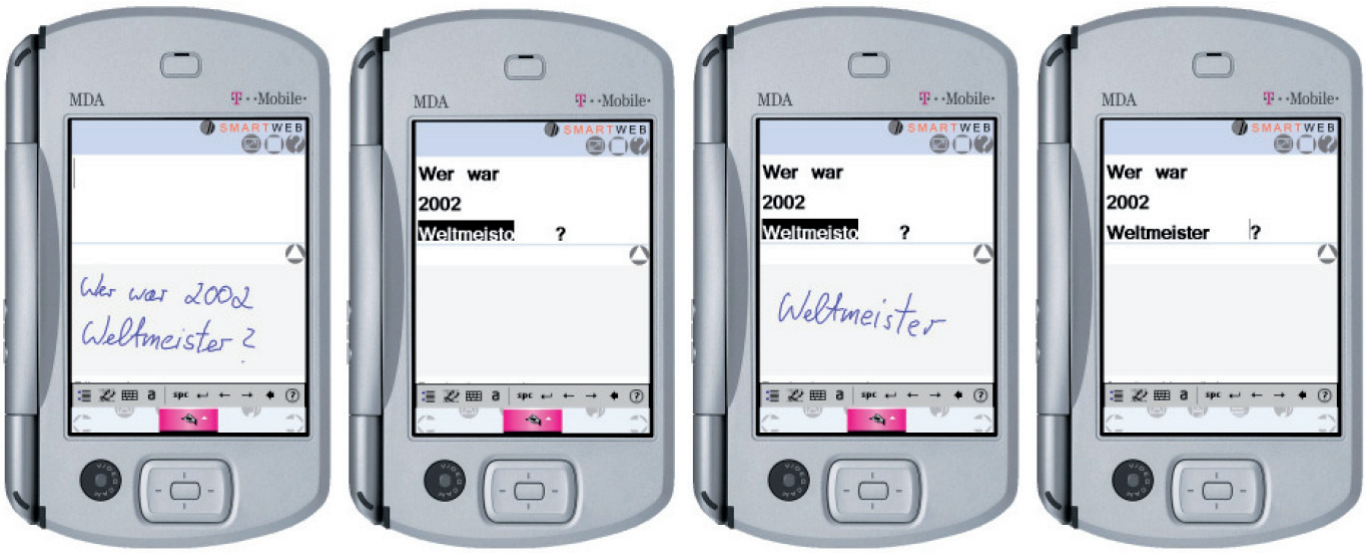
In this section we will describe the implementations of correction possibilities the user should be able to test, and the implementation regarding the presentation of multimodal results.

2.2.1 Correction Possibilities

Probably the most important question concerning the user interaction model is how to correct invalid user input which stems from recognition errors or from natural language understanding errors. New interaction methods have to be defined in order to repair an invalid user input. This becomes even more serious in the context of composite multimodality, where the dialogue system must understand and represent the multimodal input. For SmartWeb, we decided on the following correction possibilities and directed them to the physical MDA user interface.

Correction of textual query:
A textual query such as "Who was world champion in 1990?" can be edited directly in the editor text area which displays it after speech interpretation. When the editor field is clicked, the correction mode is activated and, e.g., the device keyboard, provided by the Windows mobile operating system, appears. The handwriting recogniser is preferred because it allows intuitive word selection on screen. The user could then, for example, simply click on or underline an incorrect word.

For example, the query "Wer war Weltmeister 2002 / Who was world champion in 2002?" is shown below. If the user presses the editor



button, the editor text field is enlarged. The user can switch from keyboard to transcriber before SmartWeb is started. (Tab the input panel icon at the bottom center of the screen and tab the input selector arrow. See also 'Edit text using transcriber in the Windows Mobile help page'.)

The transcriber gestures, such as quick strokes of the stylus, and options, are thoroughly available within the SmartWeb GUI. Best practice is to change to the editor mode to increase the font size when using the transcriber to correct words.

Out-of-vocabulary (OOV) word correction: The look-and-feel can then be described as follows. A click on an OOV word or word group in red font directly enables the correction mode of the word(s). The user can select other word hypotheses displayed in the pull-down menu or use the keyboard/handwriting recogniser instead. The figure on this page shows an example of out-of-vocabulary word correction.

2.2.2 Result Presentation

In this section we will discuss the presentation of results stemming from the Semantic Web access mediated by the Semantic Mediator component. It was already mentioned that visual feedback will be provided on the screen at each dialogue stage, especially in order to provide feedback for speech input. Not only user feedback, but also search results will be presented on screen at least. In addition, other modalities can be selected, but depend on the context in which the system is used and the data's suitability of being presented in a special modality.

2.2.2.1 Perceptual Feedback Requirements and Implementation

Perceptual feedback allows the user to understand the current processing state of the system, in order to react accordingly. Although we do not use a lifelike character like Smartakus (see Smartkom project), system-activity and responsiveness can be expressed by liveliness of the presentation elements in terms of metaphors. System states that are used as perceptual states are (1) *Listening/ Idle State*, (2) *Recording* (green and red icon), (3) *Understanding*, (4) *Query Processing* (status bar), (5) *Presentation Planning*, and (6) *Presenting*. Visual feedback is provided on the screen at each prominent dialogue stage, especially in order to provide feedback for speech input. Not only user feedback, but also search results will be presented on screen at least. In addition, other modalities can be selected, but depend on the context in which the system is used and the data suitability of being presented in a special modality.

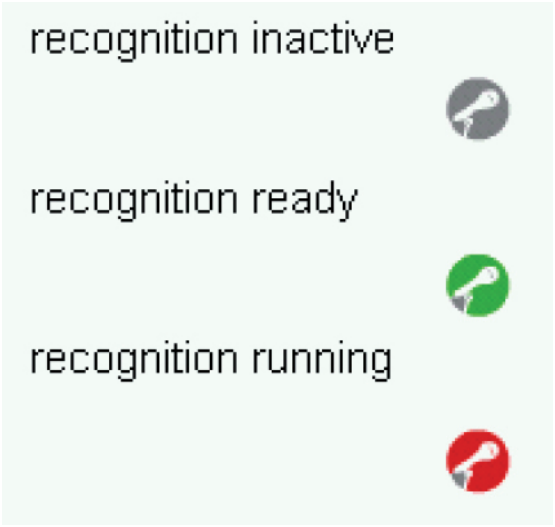
1. *Listening/Idle State:* Green micro and pulsing SmartWeb logo.

The pulsing SmartWeb logo indicates that the system is reactive towards speech, text input, and click events. When the logo is static, the system is busy. This can occur, for example, if larger XML result messages are being sent to the device. In this case, the user should wait for the logo to pulse again before addressing

SmartWeb. (The logo is also static when the progress is playing, but this is not concerned with processor load.)

The click-down area of the inactive microphone is designed to be larger than its visual appearance. This way the user can activate the micro in push-to-talk mode more easily with his thumb, holding the device in his left hand.

2. *Speech Recording/Recognition State:*
The following figure illustrates the different microphone states: recognition inactive, recognition ready, and recognition running.



3. *Understanding:* Semantic paraphrase and concept icons. Semantic paraphrases display the semantic interpretation of the user's utterance in ontology concept notation. Concept icons present feedback to question understanding and answer presenting in a language-independent, conceptual way. The movie icon corresponds to the 'Film (gefragt)/ Movie (asked)' text. The sun icon, for example, complements textual weather forecast results and conveys weather condition information.

4. *Query Processing*: Progress bar with the following emblem.

5. *Presentation Planning*: Indicated as text output in the statusbar (Resultat wird generiert / Results are being generated).

6. *Presenting:* Answers are loaded into the answer field and the media screen to be accessed by the control buttons and NaviLinks (explained further down). End of query processing is indicated in the status bar, potential result filtering is also indicated in the status bar on-time. (Recherchen abgeschlossen / Search terminated.)

Additional perceptual feedback includes:

• Zoom in selected items, implemented by *focus concept* (illustrated in the three screenshots below). The focus concept is a metaphor for the indication of a specific entity in the user's visual focus. A blue frame indicates the selection of a single table cell, which can be a graphic, an image, or a text. This information is then transferred to the dialogue engine as the focus for further processing. It helps to find references for verbal deitic utterances such as, "What is it?" Since the focus is set on one particular table cell, it is clear that the question refers to it. The focus frame is a blue transparent frame around the focus object which appears on the screen as animation. The particularity is that the other visual content apart from the focus

concept is still visible and buttons can still be pressed. The blue frame disappears if the focus is unset, or a new query is being posed, in which case the focus is unset automatically. Users can be trained to use visual foci before they utter something. This enhances the reliability of input fusion enormously.

Focus concept: Pressing a table cell sets a focus on it. If the focus is set to an image icon, the icon is focussed visually. If the focus is set to text, the text is also focussed visually. The focus concept represents a possible referent for input fusion.

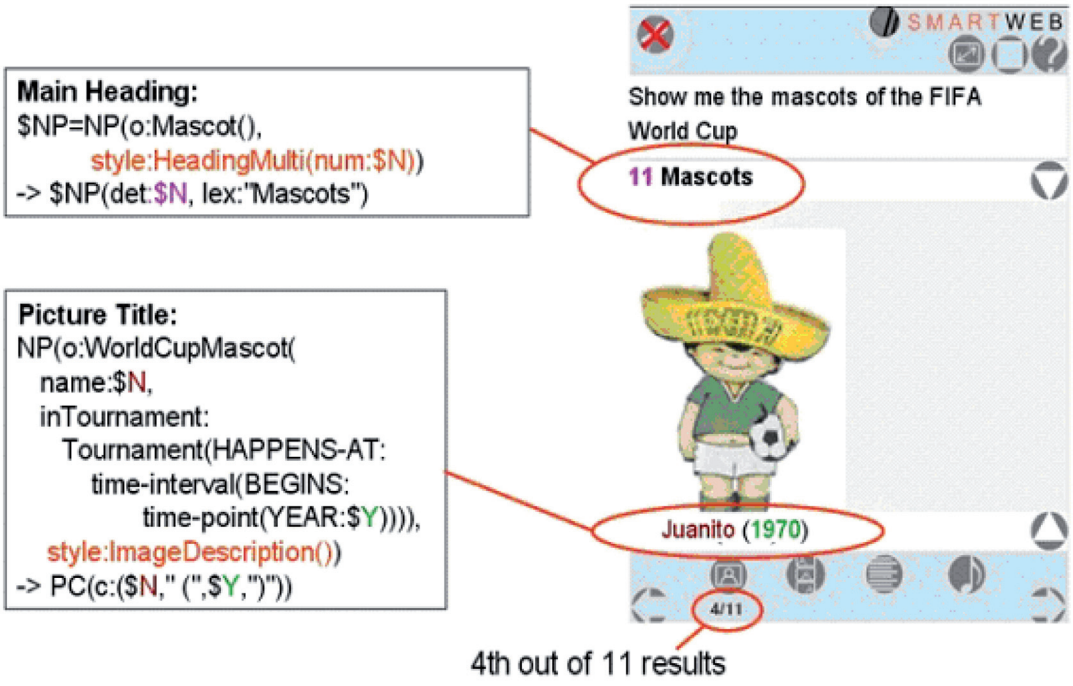
- Explanation on selection, implemented by the status bar in normal mode, and synthesis in full-screen mode. If, for example, a POI is selected, the user gets information about that POI in the status bar. The figure at the top of this page provides an example.

- Transparent popup menus: Under experimentation for textual POI information which does not fit into the status bar.

2.2.2.2 Portrait and Landscape Orientation

Portrait and landscape orientation of the same information content is shown in the two screenshots below.

In portrait orientation the user normally holds the device in one hand and uses the stylus with the other. In landscape orientation the MDA pro can also be held in both hands. The buttons in the left and right margins can be pressed with the left or right thumb, respectively.



swering sub-system re-ranks the results from the separate information extraction components, which perform passage retrieval and entity extraction as well as layout-supported extraction of images and information objects.

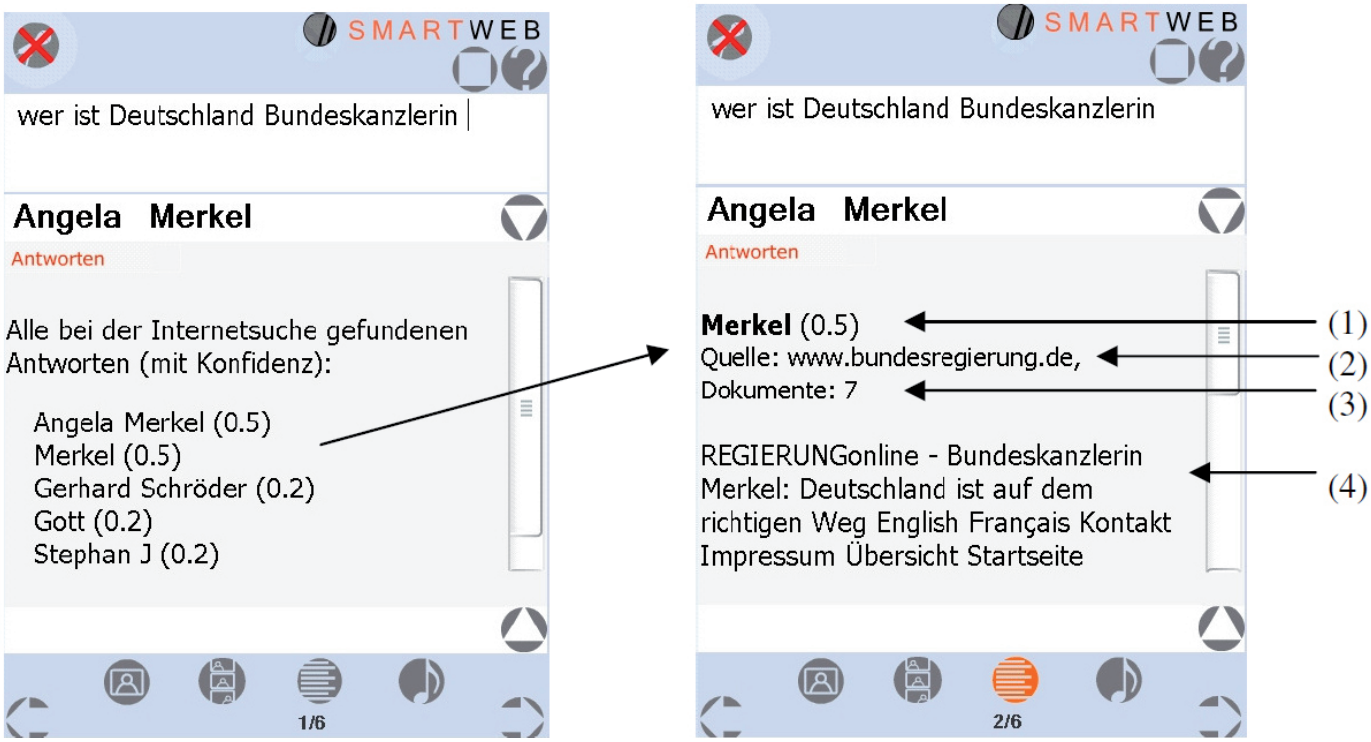
We present the five best answers to the user in the answer tab, in textual form. In addition, images can be retrieved and browsed by selecting them with a press of the image button. For each short answer, additional textual material is available when the text button is pressed; this accesses a generated answer document. Each generated text document comprises of the following information: (1) the short answer and retrieval confidence value, (2) the source server where the answer was found, (3) the document frequency of the answer within the filtered set of retrieved documents, and (4) concordance information, i.e., the paragraph where the answer has been found. One example for this is provided in the figure below.

2.2.2.3 Multimodal Answer Generation and Result Display

The input of the generation module are instances of SWIntO representing the search results. These results are then verbalized in different ways, e.g., as a heading, as a row in a table, as a picture caption title, or as text which is synthesized. Please refer to the graphic above for an example.

2.2.2.4 Result Display of Open Domain Question Answering

Open domain Question Answering retrieves answers from internet pages and PDF documents. The front-end component of the question an-



2.2.2.5 Listening to Audio

Audio material can be listened to by clicking the audio button (see image above). If the MP3 controller is activated in the preferences window, a panel with additional control buttons is displayed. A similar controller is available for video content. **Note:** Unfortunately some Linux Flash players do not support the controller button functionality.

2.2.2.6 OnFocus / OffFocus Status Display

Two components determine the attentional state of the user: the *OnView* recogniser, and the *OnTalk* recogniser. The task of the *OnView* recogniser is to determine whether the user is looking at the system or not. The *OnView* recogniser analyses a video signal captured by a video camera linked to the mobile device. For each frame, it then determines whether the user is in *OnView* or *OffView* mode. These two signals are combined to an *OnFocus/OffFocus* user state. In the *OffFocus* state, the speech input obtained from the speech recogniser is omitted in dialogue processing. The visual feedback on the screen is a blue animated microphone (see figure to the left).

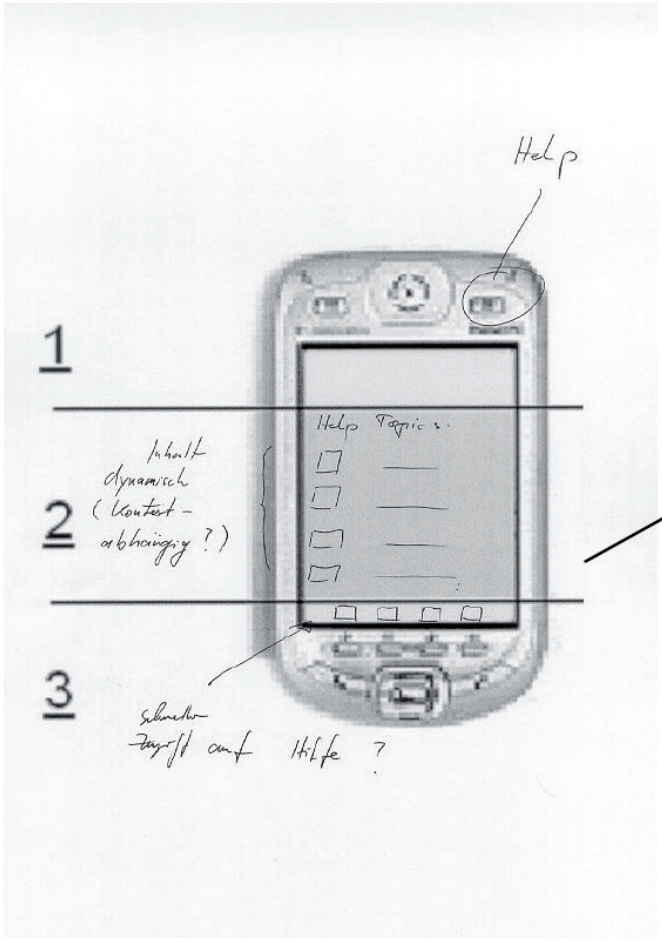
3 Interaction Metaphors Based on Storyboard

"Advances and breakthroughs in computer graphics have made visual media the basis of the modern user interface, and it is clear that graphics will play a dominant role in the way people communicate and interact with computers in the future. Indeed, as computers become more and more pervasive, and display sizes both increase and decrease, new and challenging problems arise for the effective use and generation of computer graphics.

"Recent advances in computer graphics have allowed AI researchers to integrate graphics in their systems, and on the other hand, many AI techniques have matured to the point of being easily used by non specialists. These very techniques are likely to be the vehicle by which both principles from graphic design, and the results of research into cognitive aspects of visual representations will be integrated in next generation graphical interfaces."
(Smart Graphics Enterprise, www.smartgraphics.org).

Let us start with an example showing how we implemented a special interaction metaphor in the storyboard. With this, we mean the display of help or recommendation information with the focus concept. The "Empfehlung" tab is used to present recommendations to the user. Please refer to the figure at the top right of the next page. The left figure shows the storyboard of the help information design.

A click on the help or recommendation icon enlarges the icon in order to display the type of help or recommendation. In the following example, the boy indicates a general recommendation. Special



recommendation types will be designed according to the classification of the recommended system. A click on a recommendation (in the second column of the recommendation table) activates the recommendation. This means that the recommendation is copied into the editor field. In order to send the recommendation as a request, the user can click on the cursor button (circled in red in the figure at the bottom right of the previous page). This is similar to the process of typing normal questions into the system with a keyboard.

4 More Interaction Examples

For further clarification, we will report on some tested interaction sequences. Other interaction sequences and question patterns can be downloaded at: http://smartweb.dfki.de/sys/trac/wiki/SampleUtterances.

Please make sure your dialogue server is running and you are connected. See the image to the right for a screenshot of the welcome window.

4.1 Interaction Sequence Example

Knowledge Server

1. User: “When was Germany world champion?”
2. System: “In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy),

2003 (in USA)."
3. User: "And Brazil?"
4. System: "In the following 5 years: 1958 (in Sweden), 1962 (in Chile), 1970 (in Mexico), 1994 (in USA), 2002 (in Japan)." + [*team picture, MPEG7-annotated*]
5. User: Pointing gesture on player *Aldair* + "How many goals did this player score?"

4.2 Interaction Sequence Example

Answers from Test Release 0.7, 4. March 2007:
Knowledge Server
1. User: "When was Germany world champion?"
2. System: "In the following 4 years: 1954 (in Switzerland), 1974 (in Germany), 1990 (in Italy), 2003 (in USA)."
3. User: "And Italy?"
4. System: "In the following 4 years: 1934 (in Italy), 1938 (in France), 1982 (in Spain), 2006 (in Germany)."
5. User: Click Pirlo, Andera of the 2006 World Cup + "How many goals did this player score?"
6. System: "1 goal… during the finals…"

4.3 Interaction Sequence Example

Web Services
7. User: "Where can I find Italian restaurants here?"
8. System: "Map" + Infos are displayed.
9. User: "And Greek ones?"
10. System: "Map" + Infos are displayed.
11. User: Clicks on Restaurant-POI + "How do I get here?"
12. System: "Route" + Calculated route is displayed.
13. User: "Show me ATMs."
14. System: Map + Restaurant POIs + ATM POIs are displayed.
15. User: Click on ATM-POI + "Give me more information about this."
16. System: "Bank of America" + further information, if available.

4.4 Interaction Sequence Example

Knowledge Server
17. User: "Who was world champion in 1954 ?"
18. System: "Germany"
19. User: Looks at pictures (-> Focus + deictic references are set) + "Who is the third person from the left?" ("from the top left" if there are several rows)
20. System: "Horst Eckel" + picture detail
21. User: "Go back!"
22. System: Shows answer to previous World Cup champion question again.
23. User: "Who is the player to the right of Horst Eckel?"
24. System: "Helmut Rahn."

4.5 Interaction Sequence Example

Web Services
25. User: "What will the weather be like on Saturday?"
26. System: "The weather on Saturday will be..."
27. User: "And tomorrow?"
28. System: "The weather tomorrow will be..."

4.6 Interaction Sequence Example

29. User: "How do I get from Berlin to?"
30. System: Follow-up question...

4.7 Simple Factoid Questions

"Wer war 2002 Weltmeister? / Who was world champion in 2002?" – Brasilien / Brazil (see graphic below)

4.8 More Complex Questions with More Complex Answer Structures (Answers with MPEG-7 Annotated Image Results)

"Gegen wen spielte Frankreich bei der WM 1998? / Against which teams did France play at the World Cup in 1998?" – 8 Spiele / 8 games (see graphic below)

4.9 Referencing Displayed Images

"Wann war diese Mannschaft Weltmeister?/ When was this team world champion?" – In den folgenden 5 Jahren / In the following 5 years (see graphic below)

4.10 Referencing Mpeg-7 Sub-Annotations

[player click] + "Wie viele Tore hat dieser Spieler geschossen? / How may goals did this player score?" (see graphic below)

Fullscreen and Synthesis Mode: When the user changes to fullscreen mode, picture dragging, zooming, and subselecting is possible. Dragging means the user can touch the screen and drag the picture to display a special area of the picture. The picture can be zoomed in on and out; the pointing gestures are interpreted according to the changing zoom factor. A special image part can be sub-selected by pointing on screen. If the sub-selection has a label attached to it, the label is being synthesised in the fullscreen mode, contrary to being displayed in the status/caption area in normal mode.

4.11 Referencing Mpeg-7 Sub-Annotations

[POI click] + "Wo gibt es Italiener in Saarbrücken?/ Where are Italian restaurants in Saarbruecken?" (see graphic below)

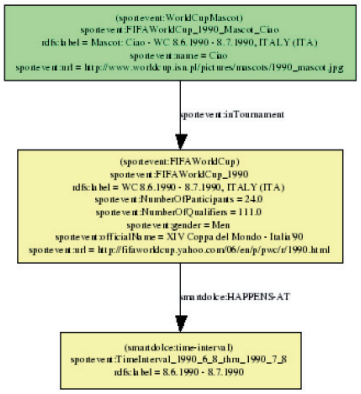
4.11 Question and Answers with Video Content

"Zeige mir ein Tor im Spiel Deutschland gegen Brasilien 2002? / Show me a goal in the 2002 match Germany against Brazil?" – 6 Tore / 6 goals (see graphic below)

Video content can be accessed by the second media button on the bottom panel. Pressing the video button, the video starts downloading and streaming.

Note: In order to enhance the playback quality we tested reducing the resolution of the GUI screen while playing. In addition, the screen listeners are reduced to two buttons, the text button and the micro button. If your version actually reduces the resolution, which can be easily seen, please press the text button after watching the video in order to return to non-video mode. In order to pose

a new question, press the micro button as usual. In any case, the video player is stopped, the resolution of the GUI screen enhanced again, and all screen listeners reactivated. (The video controller in the extra panel is still under development, please excuse the inconvenient use of the text button to return to normal mode if the resolution is changed while playing).



Note: In some environments, and due to the download speed, the video can be displayed in preferred size as soon as the download is complete. The streaming option is activated for all environments.

NaviLink structures:
In the context of new result structures in semantic ontological form (proximity graph, generated tables) as deeply nested information and knowledge structures new presentation and interaction metaphors have been developed and implemented. The NaviLink structures, which follow a kind of automatic and dynamic hyperlink generation approach, have been designed for the SmartWeb mobile handheld scenario.

"Zeige mir die Maskottchen der WM / Show me the World Cup mascots" – 11 Maskottchen/ 11 Mascots

The deeply structured results obtained from the Semantic Web access for this question can be seen in the screenshot of the OntoBroker output (http://smartweb.dfki.de/sys/svn/trunk/ontoserver/ontoserver-current-questions.html), presented at the bottom of this page.

These structures serve as input for multimodal answer generation, hyperlink generation, and presentation (please refer to the figure at the top of this page).

When asking for tables as additional inquiry **"Als Tabelle bitte/as table please"**, this part of the proximity result graph is also turned into a table structure by the generation module. This table structure is presented to the user by an additional Navilink flag "Tabelle/table". An extensive example of this is provided on the following page with a detailed explanation.

(1) In response to the question “Welche WM Maskottchen gibt es? / Which World Cup mascots are there?” the system provides a list. If the user prefers the information to be presented in a different format, he can specify “als Tabelle bitte / as table please” or type the request into the keyboard provided on screen.

(2) In our example, the user chooses to also have a table of the results which are provided in a second NaviLink tab. He can then choose to work with the results as presented initially, or switch to the provided table.

(3) When switching to the table, small icons of the mascots are provided, each of which may be clicked for a close up. The information provided in the table is modified (e.g., years are not included).

(4) In this screenshot, the user has chosen one of the mascots to look at more closely. Here we have an example of the focus concept. The focus is on the object in white with a transparent blue focus frame surrounding it.

(5) The user has chosen another mascot to examine. Again, everything except the object in focus is blue. In this combination of focus concept and table, it is possible to ask questions which refer to only the object focused on. More importantly, there is no need to make a specific reference to the object, since the focus concept has already highlighted it. Instead of asking, “What year did the World Cup have Striker as a mascot?” the user can simply ask “Which year was this the mascot?”

(6) This screenshot shows an example of the focus concept as it pertains to text only. In this case, questions may be asked specifically about World Cup Willie without the usual necessity of providing his name directly.

Cross-modal reference onto dynamically generated presentation items (Movie titles) for continued dialogue interaction: **„Was kommt heute abend in Saarbrücken im Kino? / What's playing at the cinema in Saarbruecken tonight?" – 26 Filme / 26 movies – "Wo läuft dieser Film? / Where is this movie playing?"** (see graphic above left)

„Zeige mir mehr Informationen zu diesem Film/Show me more information about this movie" (see graphic above right)

Partial queries are queries which cannot be processed at once because information is missing. These queries can only be processed by user interfaces and AI systems which either know how to supplement the query automatically or provide means for initiating a subdialogue to ask the user for missing information. According to our storyboard, we rely on dialogue manager competence to initiate clarification requests. For example: **"Wie komme ich von Saarbrücken / How do I get from Saarbruecken"** – "Rückfrage: Wohin wollen sie fahren / Follow-up request: Where do you want to go" – "**Berlin"**. Please see below for an example of how we model clarification requests on the mobile user interface.

5 Notes on GUI Elements and Implementation

Presentation elements fit their purpose best if they are not only informative, but also aesthetically satisfying; humans perceive better when the informative content is pleasing in addition to being functional. We deliver different visualisations for different data characteristics, mostly distinguished by availability of ontological result representations, Mpeg-7 annotations, and cardinality (one multimodal result, or a result list, or an ordered sequential instructive text such as a route description).

However, for the graphics and GUI surface we decided to locally play Flash-Movies on the PDA device as user interface. Flash MX 2004/Flash 8 was the latest version of Macromedia Flash in 2007. (In the context of the THESEUS project (2007-2011), we now use we now use Flash 10 for the CoMET System, and for the MEDICO Use Case we have used Adobe Flex (AiR) Technology.)

We have been mainly interested in developing interactive user interfaces quickly, and Flash is meant to provide us not only with the develop-

ment tool, but also with software for easy deployment on the mobile PDA device. During development and GUI implementation, especially for the handheld scenario, it became evident that the provided Flash running environment for mobile devices is at a very prototypical development stage. Many aspects and API functions of the desktop environment can not be used on the PDA for several special reasons, such as player version availability, execution speed, and compatibility on different platforms. Another problem is the reaction time of the GUI interface. The (touch) screen listeners operate in the Flash player plugin- and Mobile Explorer allowed time, which does not guarantee direct responsiveness. We do test all upcoming mobile Flash environments for improvements in this respect.

Due to the processor restrictions on interaction element on the PDA device, the following interaction metaphors which were planned could not, or not yet be integrated into the Handheld client distribution.

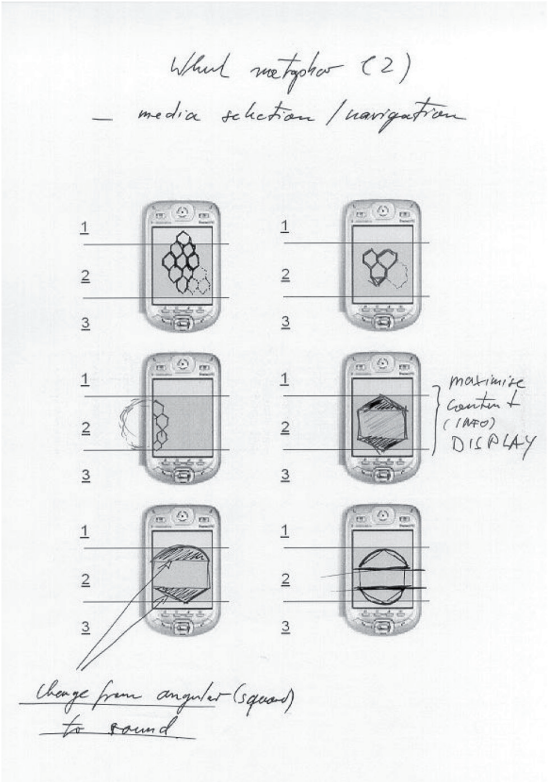
5.1 Wheel Metaphor

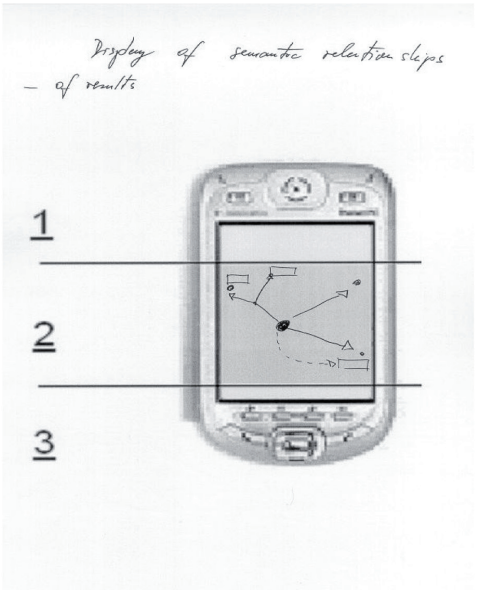
The wheel metaphor is used to generate a graphical display-selection-navigation element for long result lists, as displayed in the graphic below. Questions can be posed in natural language and assisted by pointing gestures on the screen. Natural language queries initiate an intelligent Semantic Web search. Results are presented in natural language accompanied by images, videos, text, and other media. The user can navigate through the result in an intuitive way. We recognised that the thumb plays a significant role in modern society - becoming humans' dominant haptic interactor. This development should be reflected in the interface design for future (mobile) devices. For more information, please see http://www.dfki.de/~sonntag/SMARTWEB/Interaction_Study_Haptic_Mobile_Interfaces.html.

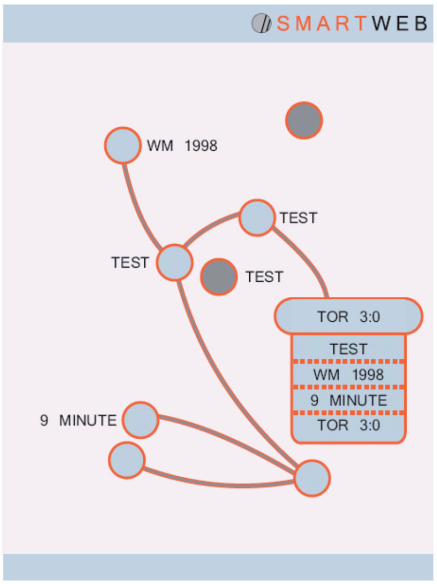
5.2 Interactive Semantic Navigation

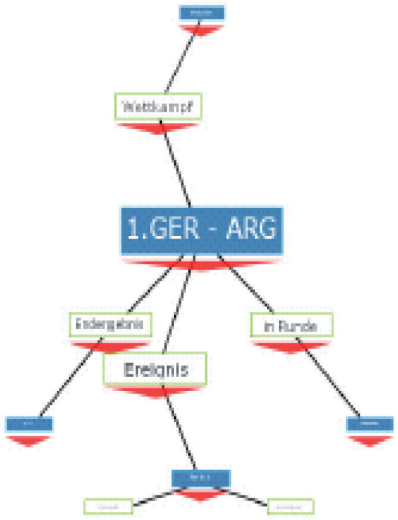
We plan to display semantic relationships of results in a graph structure. This is not a must for presenting the desired information content, but adds mostly navigational value to the presentation design. Since many of the multimodal results can be expressed in a graph-like structure, we experiment with (1) multimodal result(s) to graph conversion, and (2) automatic layouts of the derived graph structure meeting aesthetic criteria of visual presentations and elements.

Currently, several graph-layout platforms are available. We use a constraint-based general structure and layout optimiser. The graph struc-









ture and layout are produced server-side according to the solution of the layout optimiser. The resulting graph, the graph structure and positional elements will be transmitted to the device. The layouter should be started server-side on demand. In the figure above, we present the graph storyboard (left), the SmartWeb-integrated graphic design prototype (middle), and the prototype of the automatic interactive graph implementation (right). The latter uses interactive fisheye views and menus for local navigation, and constraint-based layout information for initial presentation conditions and follow-up requests.

6 Topics for Further Research

- Answering Polarisation Questions (Yes / No questions)

- Should the application look and feel like a native application? It may be easier to learn and use, but restrictions exist because of abstract commands, predefined icons, etc. For users used to the PDA, the behaviour is more predictable using native look and feel.

- Correction: (1) One step back, (2) Cancel, (3) Inline editing: Edit textual transcription of spoken input or choose other items (suggestions) by pull-down menu filled by closed classes, for example named entity classes.

- How can transaction consistency be ensured during and after back-stepping? For example, the working memory must be updated accordingly.

- Results are gathered in different languages, in German and in English. How can different language results be displayed? And synchronically?

7 References

[1] Roman Longoria, ed. Designing Software for the Mobile Context. A Practitioner's Guide, Springer (2004).

[2] Vladimir Geroimenko and Chaomei Chen. Visualizing the Semantic Web. XML-based Internet and Information Visualization, Springer (2004).

[3] R. Amant and C. Healey. Usability Guidelines for Interactive Search in Direct Manipulation Systems. North Carolina State University, Proceedings IJCAI (2001).

[4] T. Rhyme and L. Trinish, eds. Visualization Viewpoints. Perception and Painting: A search for Effective, Engaging Visualizations, IEEE (2002).

[5] R. Amant et. al. A visual interface to a music database, North Carolina State University (2002).

[6] Nicholas Cravatta, ed. The Shrinking-Interface Paradox, EDN (2003).

[7] Jaime Carbonell et. al. Vision Statement to Guide Research in Question & Answering and Text Summarization, CMU (2000).

[8] S. Salmon-Alt and L. Romary. Generating Referring Expressions in Multimodal Contexts, INLG (2000).

[9] S. Salmon-Alt. Interpreting Referring Expressions by Restructuring Context, ESSLLI (2000).

[10] N. Pfleger, J. Alexandersson, and T. Becker. A Robust and Generic Discourse Model for Multimodal Dialogue, IJCAI (2003).

[11] Elsa Pecourt and Norbert Reithinger. Multimodal Database Access on Handheld Devices. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics,* Barcelona (2004).

[12] Norbert Reithinger, Gerd Herzog, and Alassane Ndiaye. Situated Multimodal Interaction in SmartKom. In: *Computers & Graphics*, 27(6):899-903 (2003).

[13] Norbert Pfleger. Fade - An Integrated Approach to Multimodal Fusion and Discourse Processing. In: *Proceedings of the Doctoral Spotlight at ICMI 2005*, Trento, Italy (2005).

[14] Ralf Engel. Robust and Efficient Semantic Parsing of Free Word Order Languages in Spoken Dialogue Systems. In: *Proceedings of 9th Conference on Speech Communication and Technology*, Lisboa (2005).

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

[16] Norbert Reithinger and Daniel Sonntag. An Integration Framework for a Mobile Multimodal Dialogue System Accessing the Semantic Web. In: *Proceedings of Interspeech' 05*, Lisbon, Portugal (2005).

[17] Norbert Reithinger, Simon Bergweiler, Ralf Engel, Gerd Herzog, Norbert Pfeger, Massimo Romanelli, and Daniel Sonntag. A Look Under the Hood. Design and Development of the First SmartWeb System Demonstrator. In: *Proceedings of 7th International Conference on Multimodal Interfaces (ICMI 2005)*, Trento, Italy, October 04-06 (2005).

[18] Daniel 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).

[19] Daniel Sonntag and Massimo Romanelli. A Multimodal Result Ontology for Integrated Semantic Web Dialogue Applications. In: *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, May 24–26 (2006).

[20] Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfleger, Massimo Romanelli, Norbert Reithinger. SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In: *International Workshop on AI for Human Computing (AI4HC) in conjunction with (IJCAI)* (2007).

[21] Stock, O., et al.. Alfresco. Enjoying the Combination of Natural Language Processsing and Hypermedia for Information Exploration. In: *Intelligent Multimedia Interfaces,* Ed. M.T. Maybury. Menlo Park (CA): AAAI Press/The MIT Press, 197-224. (1993).

[22] Daniel Sonntag: Ontologies and Adaptivity in Dialogue for Question Answering, Heidelberg, AKA Press/IOS Press (2010).

8 Appendix

User instructions for self-running SmartWeb presentation which can be downloaded at: http://smartweb.dfki.de/Intro_Demo/start.html.

The following SmartWeb-Flash presentation has been developed to be applied in different scenarios:

1. Self-Running Presentation and Demo

The presentation consists of three sections: SmartWeb information with introductory texts, SmartWeb user interface functions, and the demonstration of answers presentation and navigation possibilities which are given to the multimodal answers of questions. For the latter, real multi-media results including videos to three sample questions are available. The user has the possibility to drag the questions into the SmartWeb user interface in order to simulate the answering of the question. The rest of the inactive interface is self-explanatory and available in both German and English with the push of a language selection button.

Pressing the space bar brings up a menu offering settings for automatic operation. The automatic operation also allows users to navigate through the questions and answers of the simulated SmartWeb system. For automatic operation, a fast content renewal (similar to a slide show) can be started, an option which is available through the menu (by pressing the space bar). Automatic operation mode (slide show) is distinctive in that the interface remains responsive. This way, an interested visitor can directly concentrate on specific areas and test their functions interactively.

When viewed in full screen mode 16:9 (or by enlarging the width of the animation), SmartWeb logos are animated in the left and right margins with the goal of attracting attention. We have decided to forgo speech syntheses of the multimodal results and other sound effects for reasons such as noise levels at the exhibition stand.

2. Interactive Documentation for New SmartWeb Users

The SmartWeb presentation offers a great amount of the documentation of the SmartWeb user interface. It interactively shows the SmartWeb handheld's interaction design and its implementation on the MDA. Additional, more technical documents are supplied for the project-wide deployment and the research community. These documents describe technical details, including the connection establishment to the dialogue system and the local message controller.

3. The Project Frame for the Evaluation of the Dialogue System and the User Interface

In the future expansion of the presentation, the running SmartWeb interface will be integrated in order to complete real evaluations of the SmartWeb (and THESEUS) dialogue system in the running dialogue sessions. The questionnaires conceived for this purpose can then be completed and saved electronically in the presentation. (This new functionality is extensively described in the book *Ontologies and Adaptivity in Dialogue-based Question Answering*.)

In respect to the evaluation of the reaction and presentation module of the dialogue manager (REAPR), the presentation should be expanded as follows. While testing the dialogue system, the user should be able to directly supply relevant feedback in order to supply empirical data for the adaptation process. (This new functionality is also described in the book *Ontologies and Adaptivity in Dialogue-based Question Answering*.)