Summative Evaluation

of the SmartWeb Prototype 1.0

Version 1.6

Hannes Mögele, Florian Schiel

Ludwig-Maximilians-Universität Munich

Technical Document No. 12
September 2007

Summative Evaluation of the SmartWeb Prototype 1.0

September 2007

Hannes Mögele, Florian Schiel

IPS LMU München
Schellingstr. 3
80799 München

Tel.: (089) 2180-5751

E-Mail: schiel@bas.uni-muenchen.de;
hannes@bas-uni-muenchen.de

This technical document belongs to sub-project 9.1:
Evaluierung

The technical document belongs to a research project that was supported with funding from the Federal Ministry of Education and Research under the funding number 01 IMD 01. The responsibility for the content lies with the authors.

Inhaltsverzeichnis

1 Introduction

This document gives a brief description of the summative evaluation that was performed on the final release of the SmartWeb prototype version 1.0 in July 2007.
The SmartWeb system is a server-based, multimodal, spoken dialogue system for accessing Internet-based resources with multimedia features. The client software runs on a mobile handheld device MDA Pro under Windows Mobile. The speech recognition, speech synthesis, query extraction, content extraction and presentation design is handled by a server application. The spoken and textual access is domain-independent; the recognition dictionary is in principle unrestricted. Connectivity between server and client is achieved by standard wireless Internet protocols such as WLAN and UMTS. Although the handheld device, the software application on client and server and the service providers form an interdependent unit the object of this evaluation is only the SmartWeb application (server and client) itself.

The technique of this evaluation is roughly based on the decisions of the Evaluation-Workshop in Munich (22. June 2006) - see the appropriate protocol - and on correspondence with interested partners, mainly DTAG. In this TechDoc we do not document the decision process for the evaluation but merely describe how the summative evaluation of the SmartWeb prototype 1.0 was realized and summarizes the results which are presented in the appendix to this document in full.

The test methodology was roughly as follows:
16 test subjects were chosen to run a series of tests up to a maximum of 6 task levels. Each level consisted of a pre-defined task that the test subject was asked to solve with the aid of SmartWeb (see section 2.2). The first test session (level 0) was designed as a training session where test subjects learned how to use SmartWeb and the MDA Pro by uttering pre-fabricated queries from a list.[1] In contrast to the previous formative evaluation ([3]) no tests with facial camera to test OnFocus detection were conducted in the summative evaluation.

The final SmartWeb prototype release 1.0 (subversion 8749) was installed in July 2007 on our server platform in Munich (see section 2.1) and tested for technical functionality (passed). Then a standard test was performed by an expert to verify that a required minimum of SmartWeb functions were available in the final release. After passing this pre-test the summative evaluation tests were performed as described in section 2.2; no technical alterations were done to the prototypes' server and client software as well as the used hardware platforms throughout the evaluation was finished.

All test conditions including meta data of test subjects (see section 2.3) as well as log files of the system and results of questionnaires were gathered in a common evaluation database (see section 2.4).

The motivation for not using one test subject per test session is based on our experiences with the SmartKom evaluation performed in the years 2003 and 2004. Naive test subjects need time to adapt to a new technology; it is not realistic to evaluate the acceptance and ergonomics of a new technology within a period of 1 hour. Also, we were interested in the general attitude of test subjects towards dialogue systems after being exposed to the SmartWeb prototypes as well as their judgment about improving technology (with each new

[1]Also, this training session enabled us to verify that the chosen test subject was no 'goat' (that is, a person where for unknown reasons the speech recognition yields only very bad results). No 'goats' were encountered during the formative evaluation.

prototype release), that is will the test subjects notice any improvements after an prototype update.

2 Basic Test Setup

In this section we briefly describe the overall design of the evaluation experiments including the technical setup, the logistics of a test series, the controlled test conditions as well as the recorded evaluation data.

2.1 Prototype Technique

Server

The server application of all SW prototype releases was installed on two Intel based hosts. The technical data of these identical hosts are:

- P4 2,80GHz
- 3GB RAM
- 100 Mbit LAN
- OS: Linux SuSE 9.2 (0.5.1) SuSE 10.1 (> 0.6)

On the first host 'host1' the SW testbed is started with the option 'services':

```
testbed.sh services
```

and after 15 sec the SW testbed is started on the second host 'host2' with option 'dialog':

```
testbed.sh dialog host1
```

'host1' is therefore the server address used by the clients.

Connectivity to the clients is realized either per UMTS (German T-Mobile) or via the in-house WLAN network using 802.1X protocol; for this purpose a driver of secureW2 was installed on all used MDA Pro clients.

Client MDA Pro

A total of three identical MDA Pro were used for the evaluation tests:

Manufacturer :	T-Mobile
Mobile type :	Twist-and-Flip
Category :	Smartphon
Networks :	GSM900/GSM1800/GSM1900/W-CDMA(UMTS)
Operating System :	Microsoft Windows Mobile 5.0
ROM Version :	ROM 1.30.113

The SmartWeb client software as well as required packages were installed using ActiveSync. No other software was allowed on the clients than the following packages:

- IBM PPRO10
- IBM PPRO10 (German)
- Microsoft .NET CF 2.0

- Macromedia Flash Player ActiveX
- SecureW2
- SmartWeb

After installation the following permanent settings were set on each client:

- AGC off
- Energy savings display: permanently on for battery, 5 min off for charging
- Energy savings system: permanently on for battery, 5 min off for charging

In the SW ConfigEditor the following permanent settings were set on each client:

- A/V: all off
- language : de
- GPS : either Berlin or Munich (depending on the task the appropriate unit was selected)
- profile / default : settings not changed

About a third of the experiments were conducted with the delivered standard T-Mobile-Headset for the MDA Pro that was cable-connected; there was no use of the Bluetooth connection. The tests were either performed indoors (office, cafe) or outdoors (cafe, open street). The outdoor tests took place under different conditions of weather and noise levels.

Tested Release

The following table shows the tested SW prototype release together with its subversion number, date of installation, pre-test result and applied operation system (OS)

Release	Subversion	Date	Pre-Test	Server OS	# of tests
1.0	8749	2007/07/09	passed	SuSE 10.1	92

2.2 Task Concept

The evaluation tests were structured into max. 6 task levels starting with a training task on level 0. Task levels are synchronized and ordered, that is, no test subject performed a task of level 3 before having performed the pervious levels 0,1 and 2. Due to the reduced project time not all test subjects completed all 6 task levels.

For each task level the test subject received a verbal instruction by the investigator as well a printed document describing the task in detail. The tasks and their general topics are:

Task Level	Topics	Environment	Number of tests
0	Training: Usage of MDA Pro, Listed Queries	office	16
1	Trip to Berlin: navigation, weather, sight seeing, hotel information	office	16

Task Level	Topics	Environment	Number of tests
2	Soccer: world series, teams, history, getting to the event	street	15
3	Cultural events, dining out, medical support	cafe	15
4	Planing a journey though Germany: geography, history, politics, celebrities	office	16
5	Journey through Germany: traffic situation, weather, sightseeing, recreation	office	14

A print-out of the original task sheets 1-5 (German) can be found in appendix 8.5.

2.3 Test Protocol

The following table shows the most prominent test conditions and their possible values for all experiments.

Condition	Possible Values	DB Key
Session Number	e[0-9][0-9][0-9]	s_session_name
Test subject code	[A-Z][A-Z][A-Z][A-Z]	s_spkcode
Date	YYYY-MM-DD	s_recordingdate
Investigator	<free text>	s_investigator
Start time	HH:MM:SS	s_start
End time	HH:MM:SS	s_end
Client	MDA_Pro_[A-E]	s_mobiletyp
Headset	MDA_Pro, none	s_headsettyp
Push-to-talk	ptt, active	s_asr_state
Environment	indoor, outdoor	s_situation
Subject is walking	yes, no	s_walk
Location	office, street, cafe	s_directions-location
Background noise	yes, no	s_background_noise
Weather	sunny, cloudy, rainy	s_weather
Age of subject	<number>	v_age
Props	<free text, e.g.'backpack', 'handbag',>	v_props
Visible piercings	yes, no	v_piercing
Subject wears Glasses	yes, no	v_glasses
Subject is bald	yes, no	v_bald_head

Condition	Possible Values	DB Key
Subject has a beard	yes, no	v_beard
Smoker	yes, no	v_smoker
Experience with dialogue systems	yes, no	v_dialog_system_experience
Experience with search engines	yes, no	v_search_engine_experience
Graduation	Hochschulreife, Mittlere Reife, Hauptschulabschluss	v_graduation
Profession	<free text>	v_profession

Test protocol data were gathered by on-screen questionnaires before and after the experiment.

2.4 Experiment Control and Evaluation Database

Each experiment was fully supervised by a trained investigator. The investigator was with the test subjects throughout the experiment. The duration of one test including preparation, instruction and interview (without postprocessing) varies between 40 – 80 minutes. The postprocessing per experiment took about 160 minutes.
Except for the initial instruction and the postprocessing all required actions of an experiment are controlled by a Perl script which calls the different questionnaires, server startup and shutdown, saving of log data etc. automatically. To simplify field tests outside the campus all control scripts may be called from any workstation in our network or even remote over an secure shell Internet connection.
All data that are gathered during the process – manually or automatically – are inserted into the Evaluation Database via remote database calls or a database web interface.

A typical experiment of the summative evaluation requires the following actions:

Preparation

- Investigator performs initial checks about server and DTAG services availability, connectivity of client (WLAN or UMTS), battery check, file space for logging.

Experiment

- Verbal instruction and handout of the task document to test subject
- Test subject fills out questionnaire A (only prior to task level 0)
- Investigator fills out test protocol questionnaire
- SW servers are started by remote secure shell and VNC; investigator checks for all SW modules working properly
- Investigator and test subject proceed to intended location
- Investigator checks for connectivity, starts SW client software and hands client and headset (if applicable) to test subject
- Test subject performs the tasks
- Investigator terminates SW client software

- Investigator and test subject return to office
- Investigator terminates SW server processes
- Log data are being saved
- Test subject fills out questionnaire B
- Test subject fills out questionnaire C (only after last test)
- Investigator handles IPR documents and payment to test subject (only after last test)

Postprocessing

- Investigator fills out test report sheet (comments, double check for important test conditions etc.)
- Client hardware is returned to lab and recharged
- Report sheet is verified to DB contents
- Recorded voice input is transcribed (see section 4)
- Interaction protocol is automatically extracted from log data (including voice input/output, presented media)
- Interaction protocol is manually segmented into interaction units and rated according a query success scheme (see section 4)
- Investigator issues money transfer as reimbursement for all performed tests (only after last test)

3 Test Subjects and Meta Data

A total of sixteen test subjects (5 male / 11 female) were carefully selected for the summative evaluation. The age ranges from 22 to 46 with an average of 26,1. Professions and education levels are solely situated in the academic environment. All 16 test subjects declared to consult Internet services on a regular basis (15/16 daily); 2 test subjects reported to use dialogue systems once a week, the others only very rarely. All test subjects have had heard speech synthesis once in a while but do not use a device with speech output (such as a navigation system or SMS reader) on a regular basis (values derived from questionnaire A, see appendix 8.7 for a complete listing of questionnaire results).
Aside from the person dependent test conditions as mentioned in section 2.3 the following additional meta data were recorded from each test subject:

Public data

Test subject code	[A-Z][A-Z][A-Z][A-Z]
Date of birth	YYYY-MM-DD
Gender	male, female
Mother tongue	<language code>
Mother tongue mother	<language code>
Mother tongue father	<language code>

Test subject code	[A-Z][A-Z][A-Z][A-Z]
Elementary School State	<state code>
Handedness	right, left, unknown
Project interest	yes, no
Comments	<free text>

Confidential data

Surname	<free text>
Name	<free text>
Street and number	<free text>
CIP code	<free text>
City	<free text>
Country	<country code>
Phone	<number>
Mobile phone	<number>
Fax	<number>
Email	<free text>

Care has been taken to separate public from confidential data. Only the database supervisor has access to the confidential part of the database.

4 Analysed Data

We distinguish three methodological different types of evaluation data: *technical, subjective* and *expert rated data*.

4.1 Technical Evaluation – Speech Recognition

As mentioned earlier all recorded voice input of the test subjects is being transcribed after the experiment by experienced phoneticians. The transcript is coded in a reduced subset of the SmartKom transliteration standard ([2]). Unnecessary tags were stripped from the transcript and different spellings of same words (e.g. 'Hauptbahnhof' vs. 'Haupt-Bahnhof'') were automatically equalized to avoid errors caused by homophones in the results.
To handle the out-of-vocabulary (OOV) problem words not to be found in the recognizer's dictionary are replaced by the tag '<OOV>'; the latter was also done with detected out-of-vocabulary parts found in the output of the speech recognizer of SW. Thus, all out-of-vocabulary words in the transcript as well as in the ASR output can be matched on each other; special classes of OOV as being produced by the recognizer are not considered.
Using a symbolic DP algorithm the resulting transcripts are automatically aligned to the

corresponding output of the speech recognition engine of SW and the word correctness and word accuracy are calculated.[1]

4.2 Subjective Evaluation - Questionnaires

The subjective evaluation aims at a general judgement of the SmartWeb prototypes, judgement of certain aspects within the communication process (such as speech recognition, media presentation, ergonomics, speech output etc.) as well as to reveal positive or negative trends during the development of the prototypes.
For this purpose three different types of questionnaires were developed in close cooperation with SW partners[2]:

- **Questionnaire A** : general attitude and experience of test subjects towards dialogue systems, PDAs, search engines, speech synthesis.
- **Questionnaire B** : detailed rating of different aspects of a test session
- **Questionnaire C** : general attitude and experience of test subjects towards dialogue systems, PDAs, search engines; quality of speech synthesis in the SmartWeb system; general features of dialogue systems

Questionnaire A was answered by each test subject **prior to the first test session**. Aside from the speaker specific meta data as described in section 3 the test subjects were asked to rate their knowledge as well as their attitude towards automatic speech interfaces. Since five of the questions have been asked after the last performed test session, the found differences may indicate a change of attitude of the test subjects towards certain aspects of Human-Machine-Interfaces caused by the exposure to the SmartWeb prototypes (see Questionnaire C).
Technically the questionnaire was presented by a GUI that was automatically called prior to the first experiment; the entered data were inserted to the Evaluation DB via remote calls.
See appendix 8.1 for a detailed description of the asked questions and appendix 8.7 for a complete listing of results.

Questionnaire B was answered by each test subject **right after each performed test session**. It comprises 29 ranking questions, two text fields to fill in positive or negative aspects of the system as well as one free text window to enter other comments. Each ranking question provided an additional 'non-applicable' button and an optional comment field; ranking widths were either from 1 ... 6 (no neutral position) or 1...5 (neutral position at 3) between two extreme rankings, e.g. '*very bad 1 2 3 4 5 very good*'. Whenever applicable the extremities were arranged in such a way that the negative extreme is positioned to the left and the positive extreme was positioned to the right. The questionnaire form was generated automatically after the session was closed using the perl script *smwe.pl* published in [1] and the inserted values were directly transferred to the Evaluation Database.
See appendix 8.2 for a detailed listing of all 29 ranking questions.

Questionnaire C was answered by each test subject **right after the last test session**. It partly contains the same questions as Questionnaire A.
The aim of this questionnaire is to detect positive / negative trends in the overall judgement / attitude towards complex spoken language interfaces after being exposed to the

[1] Correctness is defined as (N-D-S)/N while accuracy is defined as (N-D-S-I)/N where N = Number of words, S = number of substitutions, D = number of deletions and I = number of insertions.
[2] Mainly German DTAG Berlin and IMS Stuttgart

SmartWeb system. Also, the overall quality of the speech synthesis of SmartWeb as well as general features of dialogue systems are evaluated in this questionnaire.

As with all subjective evaluations the problem of the reference arises: To which expectations do the test subjects compare the performance of the system? Partly this problem can be circumvented by the precise formulation of ratings; for instance:

Erroneous inputs could easily be corrected *I do not agree ... I agree*

Other, very general questions can only be analysed in relation to earlier / later test sessions, such as:

Using SmartWeb is ... *...boring ... ...fun*

Trends can be analysed across test sessions of a single test subject in time (questionnaire B) or across test sessions of all subjects regarding the prototype (questionnaire B) or with regard of the SmartWeb experience in general (questionnaire A vs. C).

4.3 Expert Evaluation of Query Success

Each response of the SmartWeb system depends on a variety of contributing modules interacting in a complex way. Some of these modules are themselves dependent of the state of the Internet itself. It is therefore – as with all systems with a higher complexity – not possible to assess the performance of each single module in isolation and sum up all performances to achieve an overall measurable performance of the system. This does of course not imply that single modules may not be assessed alone; for instance in the technical evaluation we apply this approach to the speech recognition engine, while in the subjective evaluation special assessments about the quality of the speech output are analysed. But for other, more interconnected modules such as the dialogue manager this approach is not feasible.

We therefore adopted the traditional *End-to-End evaluation* approach successfully applied earlier in the *Verbmobil* and *SmartKom* projects to come up with a measurable value representing the *query success* solely based on input and output of the system ('black box'). Since the input and output of the SW system is multimodal and furthermore different modalities may overlap in time, this query success cannot be extracted automatically from transcripts and log data:

In a first step a so called *transaction protocol* (TP) is automatically extracted from the available log data of a test session and time aligned with the transcript of the spoken input of the test subject.
The TP is then segmented into *interaction units* (IU) where each unit must contain at least either

- one user input and the corresponding system response,
- one user input without system response, or
- one system response without a prior user input

In a second pass each IU is labelled by an human expert with regard to *task type* (TT), *input modality* (IM) and *system answer* (SA).

Possible TT values are:

Code	Task Type
SYS	System greeting at start of new test session
POI	Point of interest: queries about touristic sites
OD	Open domain: queries outside of all SW domains
W	Weather: weather forecast in cities
C	Cinema: movie titles, actual programs, contents, ...
N	Navigation: requests about directions or route plans, maps
H	Hotel information
R	Restaurant information including price lists, reservations, bars, ...
F	Soccer: Games, players, world series, locations,...
E	Events: Cultural events such as concerts, theatre, ...
HC	Health care: directions to the next physician, hospital, pharmacy, ...
SPT	Public transportation: schedules, stations, bus lines, air traffic, ...
U	Unknown: functions that are not supported by SW, OffTalk, non–cooperative behaviour
G	Geography: states, cities, populations
HI	History: history of Germany, buildings etc.
PO	Politics: parties, parliament (Reichstag)
VIP	Queries about celebrities from politics, art, sports
PG	Parking: location of garages
TS	Traffic situation (not navigation, see N)
B	Banks
+P	Picture(s) requested: including web cams, maps, etc.
+V	Video requested
+A	Audio requested
+T	Text(s) requested
+FQ	Follow–up question

For example:

Subject: "Show me a map of Munich with all the better restaurants on it."
TT: R+P
Subject: "Do you have any descriptions about these?"
TT: R+T+FQ

Possible *input modality* (IM) values are:

Code	Input Modality
S	Spoken input
T	Text input

Finally, the SA values describe the general success of the system response:

Code	System Answer
CO	Correct: SW answers the user query correctly
IC	Incorrect: SW presents an answer but the answer is not correct
PA	Partially correct: SW present several possible answers; the correct answer is among them but not at the first position
FA	Failed: SW ignores the input or recognizes the speech input but does not process it or presents the message *"Rückfrage"* and freezes
FU	Failed because of non-cooperative user behaviour, typos in text input, illogical queries, off-talk
SN	SW processes the query but reports that no answer has been found

Note that the values of SA are *independent of the success of the speech recognition engine*, that is a perfectly recognized query may still lead to a SA value of 'FA' if the system does not process the query and does not issue a negative answer. Vice versa, a partly recognized input may lead to a SA value of 'CO' if SW manages to derive the intended meaning from the corrupt recognition result.

Example: Test subject types in: *"Picture of the Sears Tower"*
The system presents a picture of the Sears Tower.
Code: TT:OD+P; IM:T; SA:CO
Test subject says: *"Do you have a picture of the top as well?"*
The system presents a picture of a top-less model.
Code: TT:OD+P+FQ; IM:S; SA:IC

5 Summary of Results

5.1 Test Conditions

The table in appendix 8.5 gives an overview about the tested conditions across the 1+5 task levels of the summative evaluation. To summarize, the following table shows the total number of tests analysed per test condition:

Total	Environment	Connetivity	Headset	ASR activation
92	68 indoor	63 WLAN	39 MDA Pro	77 PTT
	24 outdoor	29 UMTS	53 none	15 open

Test conditions are not independent. For instance there is a strong correlation between indoor/outdoor and WLAN/UMTS simply because outdoor test require an UMTS connection in most of the cases.

5.2 Technical Evaluation

The speech recognition engine of SW 1.0 was evaluated according to the method outlined in section 4.1.
The following table shows word correctness and word accuracy in percent for all conducted experiments, separated according to the acoustical environment (indoor/outdoor), the ASR activation (push-to-talk/open microphone) and the used microphone (headset vs. free speaking into the build-in microphone).[1]

Condition	all	indoor	outdoor	PTT	open	headset	build-in
Correctness	54,77%	56,52%	50,08%	54,18%	56,61%	48,84%	60,06%
Accuracy	50,92%	53,02%	45,25%	50,29%	52,90%	44,76%	56,40%
# of words	11312	8249	3063	8567	2745	5326	5986

The measured ASR correctness of 54,77% is a remarkable and significant improvement compared to the overall correctness measured during the development phase of the SmartWeb prototypes (46,59%, see [3], p. 14). However, since the task types differ between the formative and summative evaluations these values cannot be compared directly.
Indoor session scored better than outdoor sessions while there seems to be no significant difference between the conditions 'push-to-talk' and 'open microphone'. Surprisingly the usage of the headset worsened the ASR results significantly; one possible reason for this could be the selected training data for the ASR engine.

The correctness was also determined for all tests conducted by a single test subject. **Results across test subjects range from 34,0% to 70,8% correctness**[2] which indicates that there are still significant differences in the ASR performance across different speakers**.**

It is often argued that male voices score better in ASR than female voices because the male spectrum is denser packed with harmonics and can therefore express the overlaid format structure in a better way. To test this hypothesis we measured the ASR performance for both genders exclusively. As can be seen **there is in fact a significant difference between the two genders**, as being reported in literature.

[1] 100% correctness means all spoken words have been recognized correctly but there might be additional inserted words that were not actually spoken; 100% accuracy means no additional inserted words. Consequently the correctness ranges from 0 to 100% while the accuracy might become negative (if many insertions take place).
[2] With both extreme outliers being female.

Condition	all	female	male
Correctness Accuracy	54,77% 50,92%	51,82% 48,20%	61,22% 56,86%
# of words	11312	7761	3351

To see the influence of the task domain we calculated the same values for all test subjects across task types:

Task Level	Topics	Environment	Correctness / Accuracy	# of words
0	Training: usage of MDA Pro, listed queries	office	64,03% 61,66%	1265
1	Trip to Berlin: navigation, weather, sight seeing, hotel information	office	57,16% 53,62%	1809
2	Soccer: world series, teams, history, getting to the event	street	54,00% 49,14%	2039
3	Cultural events, dining out, medical support	cafe	42,10% 37,74%	1696
4	Planing a journey though Germany: geography, history, politics, celebrities	office	57,33% 53,21%	1917
5	Journey through Germany: traffic situation, weather, sightseeing, recreation	office	55,61% 52,12%	2586

As it has to be expected the ASP performance values for the training session 0 are significantly better than for the remaining task levels, since test subjects used pre-fabricated queries in task level 0. The remaining task levels are remarkable evenly distributed except for task 3. Since the ASR of SmartWeb achieved average results in the only other outdoor task 2, the reason for the worse results in task 3 are probably not caused by the acoustic environment but rather by the task itself: the topics 'cultural events, dining out, medical support' are not specific SmartWeb domains and therefore not easy to handle.

5.3 Subjective Evaluation – Questionnaires

5.3.1 Questionnaire B – Overall Judgment of SmartWeb

The following table summarizes the raw results of questionnaire B over all tests. A detailed description on the same data set together with the original question text in German is given in appendix 8.4.
The ranking scales of questionnaire B were mostly[1] arranged in a way that negative extremes are always to the left and positive to the right; therefore barycenters[2] (indicated in

[1] Exceptions are questions 10 and 21 which were exempt from the barycenter counts.
[2] Bary center defined as sum [<question-rank> * <rank-count>] / <number-of-tests>

underlined bold face) positioned in the right half of the table indicate positive ranking (15) and in the left half negative ranking (15).

		N.A.	1	2	3	4	5	6
Question:	1	0	0	2	9	13	**42**	26
Question:	2	0	0	0	0	8	38	**46**
Question:	3	8	3	8	15	**27**	19	12
Question:	4	0	1	5	9	21	**44**	12
Question:	5	3	3	14	13	**17**	**35**	7
Question:	6	0	1	14	**32**	17	23	5
Question:	7	6	12	30	**31**	8	5	0
Question:	8	0	24	**33**	**35**	0	0	0
Question:	9	0	**24**	**23**	**17**	**21**	7	0
Question:	10	0	0	5	12	**25**	**32**	18
Question:	11	1	18	**24**	9	11	**24**	5
Question:	12	0	3	6	16	19	**23**	**25**
Question:	13	0	5	**33**	**42**	12	0	-
Question:	14	0	3	18	**25**	**26**	20	-
Question:	15	0	18	**42**	25	7	0	-
Question:	16	1	3	17	21	**33**	17	-
Question:	17	0	4	29	**34**	25	0	-
Question:	18	0	**39**	**26**	14	9	4	0
Question:	19	0	20	**29**	14	14	13	2
Question:	20	0	18	**29**	11	13	17	4
Question:	21	7	3	12	13	15	**30**	12
Question:	22	0	0	3	9	7	31	**42**
Question:	23	0	23	**27**	18	16	8	0
Question:	24	0	18	**46**	25	3	0	-
Question:	25	3	**23**	**22**	11	15	12	6
Question:	26	1	1	4	2	8	28	**48**
Question:	27	1	0	8	13	**26**	**29**	15
Question:	28	0	15	**25**	**20**	20	12	0
Question:	29	0	4	8	12	20	23	**25**

It seems to be the case that the overall judgment of the Smartweb system was mostly average.
Extreme outliers are:

- + Question 2 : "Readability of text output was good" (6 of 6)
- + Question 12 : "It's easy to correct wrong inputs." (5,5 of 6)
- + Question 22 : "I always knew what to do next." (6 of 6)
- + Question 26 : "The combination of text and voice input makes sense." (6 of 6)
- - Question 18 : "Smartweb could answer my questions." (1 of 6)
- - Question 25 : "Voice input helps to speed up." (1,5 of 6)

5.3.2 Questionnaire A vs. C – Trends before/after SmartWeb

General attitude/experience of the test subjects as being asked in the first section of questionnaire A[1] have already been discussed in section 3.
Appendices 8.7 and 8.8 show the raw results from questionnaires A and C respectively.

Before/After Effects

[1] *Keys: experience_internet,experience_diasys, synthesis_**

Five questions[1] were asked in questionnaire A and C to test for before / after effects caused by the exposure to the SmartWeb system. A comparison of the results reveal that only the answers to one question differ significantly:

"How much would you be willing to pay for a service that gives you unrestricted spoken access to the Web?"

Choices	Before SmartWeb	After SmartWeb
... nothing	7	3
... 10 Cents per minute	8	11
... 25 Cents per minute	0	2
... 50 Cents per minute	1	0
... 100 Cents per minute	0	0
... more than 200 Cents per minute	0	0

In contrast to the formative evaluation (see [3], p. 19) the willingness to pay for such a service has increased by using the SmartWeb system.

Speech Synthesis
Five questions[2] have been asked in questionnaire C regarding the quality / applicability / appropriateness of the speech output. The answers reflect a moderate acceptance of the speech output with exception of the question regarding the '*naturalness of the voice'* where no test subject judged positive. Nevertheless the majority of the test subjects shared the opinion that the voice can be used in the SmartWeb system.

Special System Features
Six questions were asked in questionnaire C about certain general features of dialogue systems. Test subjects were instructed by the investigator not to give their opinion regarding the SmartWeb system but rather to dialogue systems in general. The motivation for these questions was to test whether the exposure to the SmartWeb system has heightened the awareness to certain ergonomic features.
In summary the test subjects only agreed to three important features:

- "It should always be possible to ask further questions about an already answered topic!" (15/16), e.g.:
  "*Tell me about the weather forecast in Berlin for today!" ... "What about tomorrow?"*
- "A dialogue system has always to notify the user that it is busy processing a query!" (16/16)
- "It must always be possible to correct/modify erroneous input or false speech recognition results from a previous query!" (16/16)

The majority of test subjects agreed that

- a dialogue system must acknowledge all queries right after their input
- a user should know beforehand which type of information can be retrieved using the dialogue system

Test subjects were undecided about whether the correct formulation of queries should be know beforehand to the user.

---

1 *Keys: payfor_diasys, opinion_service, opinion_human, opinion_personalassi, opinion_help*
2 *Keys: uec_voice_fit_to_system, uec_voice_quality, uec_voice_pleasentness, uec_voice_naturalness, uec_voice_applicability*

5.4 Expert Evaluation of Query Success

The log data and transcriptions of all 92 test sessions were processed and labelled according to the methodology given in section 4.3. In total 2496 interaction units were registered; from these 267 stemmed from the training session 0, which are considered separately in the following analysis.

General Query Success Rates

The following table shows the rating of the query success for all test sessions (codes explained in section 4.3). Sub-types of user induced failure (FU) are not considered:

Code	Task Levels 1–5		Task level 0	
	Count	**%**	**Count**	**%**
CO	381	17,1%	87	32,6%
IC	261	11,1%	29	10,9%
PA	53	2,4%	10	3,7%
SN	652	29,6%	52	19,5%
FA	651	29,2%	55	20,6%
FU	231	10,4%	34	12,7%
Total	2229	100,00%	267	100,00%

For an overall rating we could argue that SA types CO ('Answer correct'), PA ('Partially correct') and SN ('Processed query correctly but could not find information in the Web') should be considered as a successful task, while SA types IC ('Incorrect answer') and FA ('Failed to process query') should be considered as failures of the system. Finally, the SA type FU ('Failure because of uncooperative user behaviour') should not considered at all (for simplicity we present only the relative values in percent):

	Task Levels 1–5	Task level 0
Success (CO+PA+SN)	54,35%	63,9%
Failure (IC+FA)	45,65%	36,1%

The SmartWeb system processes every second input query correctly in the sense that it communicates a conclusive answer to the user. Not surprisingly, the training task 0 yields a significantly higher success rate.
Note that the number of cases where the system was unable to find an appropriate answer from the Web (SN) is with 29,6% still high compared to correct (17,1%) or partially correct (2,4%) answers. However, a comparison with the corresponding values from the formative evaluation (see [3], p. 20) indicates a significant improvement in terms of correct and partially correct system responses.

Task Levels 1–9	CO	IC	PA	SN	FA	FU
Formative E.	6,3%	13,7%	1,1%	35,1%	17,4%	26,4%
Summative E.	17,1%	11,1%	2,4%	29,6%	29,2%	10,4%

At the same time the proportion of successful responses which did not yield the desired information from the Web dropped from 35,1% to 29,6% and the percentage of user-induced failures decreased from 26,4% to 10,4%.
Although the values are not directly comparable since the number and type of test task differs between formative and summative evaluation, this indicates an overall performance improvement towards the final SmartWeb prototype 1.0.

Query Success vs. Input Modality

1949 IU (78,08%) have been labelled with Speech Input modality. To verify the impact of the speech recognition input we separated the query success rate as defined above according to the two different input modalities speech and text input:

Success (CO+PA+SN)	Task Levels 1–5	Task Level 0
Input Speech	50,0%	60,1%
Input Text	69,6%	80,5%

The text input modality yields a significantly higher query success rate for both, the training task level as well as the remaining task levels as it is to be expected, since the speech input engine often delivers wrong or incomplete input to the system.

Query Success vs. Task Type

Since the SmartWeb system handles different types of queries in a different manner, we analysed the query success rates as defined above for the different task types as defined in section 4.3. The following table shows the overall occurrence of a task type in the task levels 1-5 and the relative query success within all task types. Statistically valid values are set in bold face; the occurrence of the remaining task types was too low to yield reliable results.

Code	Task Type	Occurrence	Success (CO+PA+SN) Task Levels 1–5
POI	Point of interest: queries about touristic sites	**14,3%**	**42,9%**
OD	Open domain: queries outside of all SW domains	2,71%	70,1%
W	Weather: weather forecast in cities	**8,5%**	**65,7%**
C	Cinema: movie titles, actual programs, contents, ...	**9,2%**	**52,6%**
N	Navigation: requests about directions or route plans, maps	**7,1%**	**68,5%**

Code	Task Type	Occurrence	Success (CO+PA+SN) Task Levels 1–5
H	Hotel information	4,7%	45,6%
R	Restaurant information including price lists, reservations, bars, ...	4,9%	68,6%
F	Soccer: Games, players, world series, locations,...	**15,1%**	**57,6%**
E	Events: Cultural events such as concerts, theatre, ...	1,1%	70,0%
HC	Health care: directions to the next physician, hospital, pharmacy, ...	3,7%	51,2%
SPT	Public transportation: schedules, stations, bus lines, air traffic, ...	0,5%	–
G	Geography: states, cities, populations	4,6%	51%
HI	History: history of Germany, buildings etc.	3,8%	29,6%
PO	Politics: parties, parliament (Reichstag)	3,4%	61,5%
VIP	Queries about celebrities from politics, art, sports	**7,4%**	**44,3%**
PG	Parking: location of garages	1,4%	–
TS	Traffic situation (not navigation, see N)		
B	Banks	1,6%	–
U[1]	Unknown: functions that are not supported by SW, Off–Talk, non–cooperative behaviour	–	–

The query success appears to be quite uniform across the SmartWeb task types.

[1] Task type 'Unknown' was in the majority classified with SA = FU, that is a 'user induced failure' and can therefore not be considered for the query success as defined above.

6 Conclusion

The technical evaluation of the speech recognition with respect to *word correctness* yields an average value of 54,8% without any significant differences between the test conditions indoor/outdoor, PTT/open microphone and headset/build-in microphone.
The measured *word accuracy* is about 50,9% which indicates that the number of insertion errors is in a moderate range.
The *word correctness* is remarkable evenly distributed across different task domains; tasks which the SW system was designed for (e.g. Soccer World Series 2006) did not score better than tasks with completely open topics. The only exception was the outdoor task ('cafe') with topics 'cultural events, dining out, medical support', which were not specific SmartWeb domains. Although not directly comparable to the measurements during the formative evaluation ([3]) these results indicate significantly improved values in terms of speech recognition.
Considering the fact that we conducted a field trial under realistic conditions (outdoors, heavy noise, real network) and the fact that the system is able to process incomplete or malformed input queries this is a satisfying result.

The analysis of the test subjects' questionnaire shows that the overall judgment of the Smartweb system was mostly average with a slight tendency to positive judgments. There is no significant change compared to the formative evaluation.

The expert rating of 2496 interaction units with regard to the query success shows that the SmartWeb system processes 54% of the well-formed input queries correctly in the sense that it communicates a conclusive answer to the user. Note that the number of cases where the system was nevertheless unable to find an appropriate answer from the Web is with 29,6% still high compared to correct (17,1%) or partially correct (2,4%) answers. Although it should be noted here that this ratio improved significantly against the findings during the formative evaluations.
Query success rates are significantly higher for text input modality (69%) than speech input modality (50%) which has to be expected. However, 78% of input queries were in fact performed in speech which implies that the test subjects nevertheless preferred spoken input over the more reliable but tedious text input modality.
A break-up of the measured success rates across task types showed a rather even distribution for task types that occurred frequently enough to be statistically relevant. Considering the fact that the evaluation tasks contained topics from dedicated Smartweb domains as well as totally unknown domains this indicates that the SmartWeb system is surprisingly robust against unknown topics.

7 References

[1] Chin, J.P., Diehl, V.A., Norman, K.L. (1988) Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. ACM CHI'88 Proceedings, pp. 213-218.

[2] D. Oppermann, S. Burger, S. Rabold, N. Beringer: Transliteration spontansprachlicher Daten – Transliterationslexikon SmartKom. SmartKom TechDok 2 – 2001.

[3] H. Mögele, F. Schiel (2007) Formative Evaluation of the SmartWeb Prototypes. SmartWeb TechDok 11 - 2007.

8 Appendix

8.1 Questionnaire A

The following questions have been asked:

Key: uea_experience_internet
Question=**"Ich rufe Informationen aus dem Internet ab und zwar..."**
Pulldown: **mehrmals am Tag**
etwa einmal pro Tag
ein paar Mal pro Woche
höchstens ein paar Mal im Monat
praktisch nie

Key: uea_experience_diasys
Question=**"Wie häufig verwenden Sie automatische Dialogsysteme? Z.B. telefonische Banksysteme, Autosteuerung per Sprache, autom. Auskunft"**
Pulldown: **Ich entwickele selber Dialogsysteme**
Ich verwende sie jeden Tag
Ich verwende sie jede Woche
Ich verwende sie nicht mehr als einmal im Monat
Ich verwende nur selten ein Dialogsystem
Ich habe noch nie ein Dialogsystem benutzt

Key: uea_pre_opinion_diasys
Question=**"Angenommen es gibt ein perfekt funktionierendes Auskunftsystem, dann spreche ich..."**
Pulldown: **1 ... trotzdem lieber mit einem Menschen**
2
3
4
5 ... lieber mit der Maschine

Key: uea_pre_payfor_diasys
Question=**"Für einen Service, der mir mit alltäglicher Sprache unbegrenzten Zugang zu Web-Inhalten bietet, wäre ich ..."**
Pulldown: **nicht bereit, etwas zu zahlen**
bereit, 10 Cent pro Minute zu zahlen
bereit, 25 Cent pro Minute zu zahlen
bereit, 50 Cent pro Minute zu zahlen
bereit, 1 Euro pro Minute zu zahlen
mehr als 2 Euro pro Minute zu zahlen

Key: uea_pre_opinion_service
Question=**"Ein intelligentes Dialogsystem kann mir niemals den gleichen Service bieten wie ein Mensch."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch

Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uea_pre_opinion_human
Question=**"Ich lege neben der reinen Information auch großen Wert auf die menschliche Seite der Kommunkation."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uea_pre_opinion_personalassi
Question=**"Ich hätte gerne einen persönlichen Assistenten (Mensch oder Maschine), mit dem ich zu jeder Zeit auf natürliche Weise auf Inhalte des Webs zugreifen kann."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uea_synthesis_simple
Question=**"Wie oft haben Sie schon eine einfache automatische Telefonansage gehört?"**
Pulldown: **sehr oft**
oft
einige Male
selten
nie

Key: uea_synthesis_sms
Question=**"Wie oft haben Sie schon einen SMS-Vorlese-Service gehört?"**
Pulldown: **sehr oft**
oft
einige Male
selten
nie

Key: uea_synthesis_nav_simple
Question=**"Wie oft haben Sie schon die Stimme eines Nav.systems ohne Ansage von Straßen- und Ortsnamen gehört?"**
Pulldown: **sehr oft**
oft
einige Male
selten
nie

Key: uea_synthesis_nav_complex

Question=**"Wie oft haben Sie schon die Stimme eines Nav.systems mit Ansage von Strassen- und Ortsnamen gehört?"**
Pulldown: **sehr oft**
oft
einige Male
selten
nie

Key: uea_pre_opinion_help
Question=**"Ein Auskunftsystem muss in der Lage sein, dem Benutzer zu helfen, wie er am schnellsten an die gesuchte Information kommen kann."**
Pulldown: **Finde ich extrem wichtig**
Ziemlich wichtig
Tendenziell wichtig
Tendenziell unwichtig
Eher unwichtig
Halte ich für völlig irrelevant

8.2 Questionnaire B

A screen shot of the Web form is shown in fig. A1.
The following ranking questions have been asked (ranking width in brackets):

1 Die Bedienung von SmartWeb ist ...
schwierig ... leicht (6)
2 Die Lesbarkeit der Schrift war...
sehr schlecht ... sehr gut (6)
3 Die Hervorhebungen erleichterten die Bedienung.
stimme ich nicht zu ... stimme ich zu (6)
4 Die Anordnung der Informationen auf dem Display finde ich...
verwirrend ... übersichtlich (6)
5 Die Ausdrucksweise von SmartWeb war...
inkonsistent ... konsistent (6)
6 SmartWeb zeigt mir, wenn es beschäftigt ist.
nie ... immer (6)
7 Die Fehlermeldungen von SmartWeb sind...
wenig hilfreich ... sehr hilfreich (6)
8 Die Geschwindigkeit von SmartWeb fand ich...
zu langsam ... zu schnell (5)
9 Das SmartWeb funktionierte...
unzuverlässig ... reibungslos (6)
10 SmartWeb lieferte...
zu viele Informationen ... zu wenig Informationen (6)
11 So wie SmartWeb in der heutigen Sitzung funktioniert hat, könnte auch jeder andere damit umgehen.
stimme ich nicht zu ... stimme ich zu (6)
12 Fehleingaben konnte ich leicht korrigieren.
stimme ich nicht zu ... stimme ich zu (6)
13 Wie angenehm fanden Sie die Stimme?
sehr unangenehm ... sehr angenehm (5)

14 Welche Anstrengung war nötig, um die Äußerungen zu verstehen?
selbst größte Anstrengung reicht nicht zum Verstehen ... es war keine Anstrengung zum Verstehen erforderlich (5)
15 Wie würden Sie die Natürlichkeit der Stimme einschätzen?
sehr unnatürlich ... sehr natürlich (5)
16 Fanden Sie heute bestimmte Wörter schwer zu verstehen?
ständig ... nie (5)
17 Wie würden Sie insgesamt die Sprachqualität der gehörten Äußerungen beurteilen?
sehr schlecht ... sehr gut (5)
18 Die Aktivierung der Spracheingabe fand ich ...
einfach ... umständlich (6)
19 SmartWeb hat mich schnell zur gewünschten Information geführt.
trifft nicht zu trifft zu (6)
20 SmartWeb hat mir die richtigen Informationen geliefert.
trifft nicht zu trifft zu (6)
21 Der Gesprächsverlauf wurde eher von...
SmartWeb bestimmt ... mir selber bestimmt (6)
22 Ich wusste immer, wie ich die nächste Eingabe (per Sprache, Tastatur, Stift) machen konnte.
trifft nicht zu ... trifft zu (6)
23 Ich musste meine Eingaben wiederholen.
praktisch jedes mal ... nie (6)
24 Die Pausen zwischen Eingabe und Antwort erschienen mir ...
sehr lang ... sehr kurz (5)
25 Mit Hilfe der Spracheingabe komme ich schneller ans Ziel
trifft nicht zu ... trifft zu (6)
26 Die Kombination von Stift- und Spracheingabe finde ich sinnvoll
trifft nicht zu ... trifft zu (6)
27 Im Gespräch mit SmartWeb fühlte ich mich ...
unwohl ... wohl (6)
28 Ich bin von der Leistung von SmartWeb ...
enttäuscht ... beeindruckt (6)
29 Der Umgang mit SmartWeb hat ...
mich gelangweilt ... mir Spaß gemacht (6)

SmartWeb Evaluation Questionnaire

Mit diesem Fragebogen beurteilen Sie die soeben erfolgte Benutzung des SmartWeb Systems. Das Ausfüllen dauert etwa 15 Minuten.

- Bitte beantworten Sie unbedingt alle Fragen.
- Bei Fragen, die Ihrer Meinung nach nicht auf das soeben durchgeführte Experiment passen, wählen Sie bitte den Knopf: **NA**
- Sie können bei Bedarf einzelne Fragen durch Klicken auf das Icon kommentieren
- Erst wenn Sie alle Fragen beantwortet und etwaige Kommentare abgegeben haben, klicken Sie zum Abschluss auf: **Daten sichern**

Sprecherkürzel: **Nachname:** **Session-ID:**

Insgesamte Beurteilung								**NA**
1. Die Bedienung von SmartWeb ist ...	schwierig	○	○	○	○	○	○ leicht	○
Der Bildschirm von SmartWeb								**NA**
2. Die Lesbarkeit der Schrift war...	sehr schlecht	○	○	○	○	○	○ sehr gut	○
3. Die Hervorhebungen erleichterten die Bedienung.	stimme ich nicht zu	○	○	○	○	○	○ stimme ich zu	○
4. Die Anordnung der Informationen auf dem Display finde ich...	verwirrend	○	○	○	○	○	○ übersichtlich	○
Ausgaben von SmartWeb								**NA**
5. Die Ausdrucksweise von SmartWeb war...	inkonsistent	○	○	○	○	○	○ konsistent	○
6. SmartWeb zeigt mir, wenn es beschäftigt ist.	nie	○	○	○	○	○	○ immer	○
7. Die Fehlermeldungen von SmartWeb sind...	wenig hilfreich	○	○	○	○	○	○ sehr hilfreich	○
Fähigkeiten von SmartWeb								**NA**
8. Die Geschwindigkeit von SmartWeb fand ich...	zu langsam	○	○	○	○	○	○ zu schnell	○
9. Das SmartWeb funktionierte...	unzuverlässig	○	○	○	○	○	○ reibungslos	○
10. SmartWeb lieferte...	zu viele Informationen	○	○	○	○	○	○ zu wenig Informationen	○
11. So wie SmartWeb in der heutigen Sitzung funktioniert hat, könnte auch jeder andere damit umgehen.	stimme ich nicht zu	○	○	○	○	○	○ stimme ich zu	○
12. Fehleingaben konnte ich leicht korrigieren.	stimme ich nicht zu	○	○	○	○	○	○ stimme ich zu	○
Die folgenden Fragen beziehen sich auf die synthetische Stimme in der *heutigen Sitzung*, nicht auf frühere Sitzungen								**NA**
13. Wie angenehm fanden Sie die Stimme?	sehr unangenehm	○	○	○	○	○	○ sehr angenehm	○
14. Welche Anstrengung war nötig, um die Äußerungen zu verstehen?	selbst größte Anstrengung reicht nicht zum Verstehen	○	○	○	○	○	○ es war keine Anstrengung zum Verstehen erforderlich	○
15. Wie würden Sie die Natürlichkeit der Stimme einschätzen?	sehr unnatürlich	○	○	○	○	○	○ sehr natürlich	○

Fig. A1 : Screen shot of questionnaire B

8.3 Questionnaire C

The following questions have been asked:

Key: uec_post_payfor_diasys
Question: **"Für einen Service, der mir jederzeit auf natürliche Weise unbegrenzten Zugang zu Web-Inhalten bietet, wäre ich ..."**
Pulldown: **nicht bereit, etwas zu zahlen**
bereit, 10 Cent pro Minute zu zahlen
bereit, 25 Cent pro Minute zu zahlen
bereit, 50 Cent pro Minute zu zahlen
bereit, 1 Euro pro Minute zu zahlen
mehr als 2 Euro pro Minute zu zahlen

Key: uec_post_opinion_service
Question: **"Ein intelligentes Dialogsystem kann mir niemals den gleichen Service bieten wie ein Mensch."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig

Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_post_opinion_human
Question: **"Ich lege neben der reinen Information auch großen Wert auf die menschliche Seite der Kommunkation."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_post_opinion_personalassi
Question: **"Ich hätte gerne einen persönlichen Assistenten, mit dem ich zu jeder Zeit auf natürliche Weise auf Inhalte des Webs zugreifen kann."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_post_opinion_help
Question: **"Ein Auskunftsystem muss in der Lage sein, dem Benutzer zu helfen, wie er am schnellsten an die gesuchte Information kommen kann."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_voice_fit_to_system
Question: **"Wie gut passt die Stimme zum getesteten System?"**
Pulldown: **Sehr gut**
Gut
Ordentlich
Schlecht
Sehr schlecht

Key: uec_voice_quality
Question: **"Wie würden Sie nun, nachdem Sie mehrere Sitzungen mit SmartWeb absolviert haben, insgesamt die Sprachqualität der Systemäußerungen beurteilen?"**
Pulldown: **Sehr gut**
Gut
Ordentlich
Schlecht
Sehr schlecht

Key: uec_voice_pleasentness
Question: **"Wie angenehm fanden Sie die Stimme insgesamt, in allen Sitzungen?"**
Pulldown: **Sehr angenehm**
Angenehm
Neutral
Unangenehm
Sehr unangenehm

Key: uec_voice_naturalness
Question: **"Wie würden Sie die Natürlichkeit der Stimme insgesamt, in allen Sitzungen, einschätzen?"**
Pulldown: **Sehr natürlich**
Natürlich
Neutral
Unnatürlich
Sehr unnatürlich

Key: uec_voice_applicability
Question: **"Finden Sie, dass die Stimme im SmartWeb-System eingesetzt werden kann?"**
Pulldown: **Ja**
Nein

Anleitung für Versuchsleiter:
Ab hier beziehen sich die Fragen bzw. Aussagen ganz allgemein auf Intelligente Dialogsysteme und nicht notwendigerweise auf SmartWeb D.h. es soll nicht geprüft werden, ob SmartWeb diese Egenschaft hat, sondern, ob so eine Eigenschaft in einem System wie SmartWeb wünschenswert wäre oder nicht.

Key: uec_additional_questions
Question: **"Es sollte möglich sein, zu einer erhaltenen Information Zusatzfragen stellen zu können."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_defined_formulation
Question: **"Man sollte von Anfang an wissen, wie man seine Fragen formulieren soll."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_know_infotype

Question: **"Man sollte von Anfang an wissen, welche Art von Informationen ein System liefern kann."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_acknowledge
Question: **"Ein gut funktionierendes System sollte jede meiner Eingaben bestätigen."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_processing
Question: **"Ein gut funktionierendes System zeigt mir an, wenn es beschäftigt ist."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

Key: uec_modify_input
Question: **"Falsche Eingabe oder falsch erkannte Eingaben sollten leicht korrigiert werden können."**
Pulldown: **Stimme ich voll zu**
Weitgehend richtig
Tendenziell richtig
Tendenziell falsch
Weitgehend falsch
Stimme ich überhaupt nicht zu

8.4 Raw Results of Summative Evaluation – Questionnaire B

```
Es wird jeweils die laufende Nummer der Fragen (Frage 1), das Item
('Die Bedienung von SmartWeb ist ...') mit Attributen ('schwierig' - 'leicht')
gegeben. Die Ziffer zwischen den Attributen gibt die Skalierung ('[5]' oder
'[6]') wieder.

================================================================================
Frage 1
--------------------------------------------------------------------------------
Die Bedienung von SmartWeb ist ...
schwierig [6] leicht
----------------------------
| 1 | 2 | 3 | 4  | 5  | 6  |
|---|---|---|----|----|----|
| 0 | 2 | 9 | 13 | 42 | 26 |
```

```
---------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 2
--------------------------------------------------------------------------------
Die Lesbarkeit der Schrift war...
sehr schlecht [6] sehr gut
---------------------------
| 1 | 2 | 3 | 4 | 5  | 6  |
|---|---|---|---|----|----|
| 0 | 0 | 0 | 8 | 38 | 46 |
---------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 3
--------------------------------------------------------------------------------
Die Hervorhebungen erleichterten die Bedienung.
stimme ich nicht zu [6] stimme ich zu
-----------------------------
| 1 | 2 | 3  | 4  | 5  | 6  |
|---|---|----|----|----|----|
| 3 | 8 | 15 | 27 | 19 | 12 |
-----------------------------
keine Wertung: 8
================================================================================

================================================================================
Frage 4
--------------------------------------------------------------------------------
Die Anordnung der Informationen auf dem Display finde ich...
verwirrend [6] übersichtlich
-----------------------------
| 1 | 2 | 3 | 4  | 5  | 6  |
|---|---|---|----|----|----|
| 1 | 5 | 9 | 21 | 44 | 12 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 5
--------------------------------------------------------------------------------
Die Ausdrucksweise von SmartWeb war...
inkonsistent [6] konsistent
-----------------------------
| 1 | 2  | 3  | 4  | 5  | 6 |
|---|----|----|----|----|---|
| 3 | 14 | 13 | 17 | 35 | 7 |
-----------------------------
keine Wertung: 3
================================================================================

================================================================================
Frage 6
--------------------------------------------------------------------------------
SmartWeb zeigt mir, wenn es beschäftigt ist.
nie [6] immer
-----------------------------
| 1 | 2  | 3  | 4  | 5  | 6 |
|---|----|----|----|----|---|
| 1 | 14 | 32 | 17 | 23 | 5 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 7
```

```
--------------------------------------------------------------------------------
Die Fehlermeldungen von SmartWeb sind...
wenig hilfreich [6] sehr hilfreich
----------------------------
| 1  | 2  | 3  | 4 | 5 | 6 |
|----|----|----|---|---|---|
| 12 | 30 | 31 | 8 | 5 | 0 |
----------------------------
keine Wertung: 6
================================================================================

================================================================================
Frage 8
--------------------------------------------------------------------------------
Die Geschwindigkeit von SmartWeb fand ich...
zu langsam [5] genau richtig
----------------------------
| 1  | 2  | 3  | 4 | 5 | 6 |
|----|----|----|---|---|---|
| 24 | 33 | 35 | 0 | 0 | 0 |
----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 9
--------------------------------------------------------------------------------
Das SmartWeb funktionierte...
unzuverlässig [6] reibungslos
----------------------------
| 1  | 2  | 3  | 4  | 5 | 6 |
|----|----|----|----|---|---|
| 24 | 23 | 17 | 21 | 7 | 0 |
----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 10
--------------------------------------------------------------------------------
SmartWeb lieferte...
zu viele Informationen [6] zu wenig Informationen
----------------------------
| 1 | 2 | 3  | 4  | 5  | 6  |
|---|---|----|----|----|----|
| 0 | 5 | 12 | 25 | 32 | 18 |
----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 11
--------------------------------------------------------------------------------
So wie SmartWeb in der heutigen Sitzung funktioniert hat, könnte auch jeder
andere damit umgehen.
stimme ich nicht zu [6] stimme ich zu
----------------------------
| 1  | 2  | 3 | 4  | 5  | 6 |
|----|----|---|----|----|---|
| 18 | 24 | 9 | 11 | 24 | 5 |
----------------------------
keine Wertung: 1
================================================================================

================================================================================
Frage 12
--------------------------------------------------------------------------------
Fehleingaben konnte ich leicht korrigieren.
stimme ich nicht zu [6] stimme ich zu
----------------------------
| 1 | 2 | 3  | 4  | 5  | 6  |
```

```
|---|---|----|----|----|----|
| 3 | 6 | 16 | 19 | 23 | 25 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 13
--------------------------------------------------------------------------------
Wie angenehm fanden Sie die Stimme?
sehr unangenehm [5] sehr angenehm
-----------------------------
| 1 | 2  | 3  | 4  | 5 | 6 |
|---|----|----|----|---|---|
| 5 | 33 | 42 | 12 | 0 | 0 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 14
--------------------------------------------------------------------------------
Welche Anstrengung war nötig, um die Äußerungen zu verstehen?
selbst größte Anstrengung reicht nicht zum Verstehen [5] es war keine
Anstrengung zum Verstehen erforderlich
-----------------------------
| 1 | 2  | 3  | 4  | 5  | 6 |
|---|----|----|----|----|---|
| 3 | 18 | 25 | 26 | 20 | 0 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 15
--------------------------------------------------------------------------------
Wie würden Sie die Natürlichkeit der Stimme einschätzen?
sehr unnatürlich [5] sehr natürlich
-----------------------------
| 1  | 2  | 3  | 4 | 5 | 6 |
|----|----|----|---|---|---|
| 18 | 42 | 25 | 7 | 0 | 0 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 16
--------------------------------------------------------------------------------
Fanden Sie heute bestimmte Wörter schwer zu verstehen?
ständig [5] nie
-----------------------------
| 1 | 2  | 3  | 4  | 5  | 6 |
|---|----|----|----|----|---|
| 3 | 17 | 21 | 33 | 17 | 0 |
-----------------------------
keine Wertung: 1
================================================================================

================================================================================
Frage 17
--------------------------------------------------------------------------------
Wie würden Sie insgesamt die Sprachqualität der gehörten Äußerungen beurteilen?
sehr schlecht [5] sehr gut
-----------------------------
| 1 | 2  | 3  | 4  | 5 | 6 |
|---|----|----|----|---|---|
| 4 | 29 | 34 | 25 | 0 | 0 |
-----------------------------
keine Wertung: 0
================================================================================
```

```
================================================================================
Frage 18
--------------------------------------------------------------------------------
SmartWeb konnte meine Fragen beantworten.
keine einzige [6] praktisch alle
----------------------------
| 1  | 2  | 3  | 4 | 5 | 6 |
|----|----|----|---|---|---|
| 39 | 26 | 14 | 9 | 4 | 0 |
----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 19
--------------------------------------------------------------------------------
SmartWeb hat mich schnell zur gewünschten Information geführt.
trifft nicht zu [6] trifft zu
-----------------------------
| 1  | 2  | 3  | 4  | 5  | 6 |
|----|----|----|----|----|---|
| 20 | 29 | 14 | 14 | 13 | 2 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 20
--------------------------------------------------------------------------------
SmartWeb hat mir die richtigen Informationen geliefert.
trifft nicht zu [6] trifft zu
-----------------------------
| 1  | 2  | 3  | 4  | 5  | 6 |
|----|----|----|----|----|---|
| 18 | 29 | 11 | 13 | 17 | 4 |
-----------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 21
--------------------------------------------------------------------------------
Der Gesprächsverlauf wurde eher von...
SmartWeb bestimmt [6] mir selber bestimmt
-----------------------------
| 1 | 2  | 3  | 4  | 5  | 6  |
|---|----|----|----|----|----|
| 3 | 12 | 13 | 15 | 30 | 12 |
-----------------------------
keine Wertung: 7
================================================================================

================================================================================
Frage 22
--------------------------------------------------------------------------------
Ich wusste immer, wie ich die nächste Eingabe (per Sprache, Tastatur, Stift)
machen konnte.
trifft nicht zu [6] trifft zu
---------------------------
| 1 | 2 | 3 | 4 | 5  | 6  |
|---|---|---|---|----|----|
| 0 | 3 | 9 | 7 | 31 | 42 |
---------------------------
keine Wertung: 0
================================================================================

================================================================================
Frage 23
--------------------------------------------------------------------------------
Ich musste meine Eingaben wiederholen.
```

praktisch jedes mal [6] nie

1	2	3	4	5	6
23	27	18	16	8	0

keine Wertung: 0

Frage 24

Die Pausen zwischen Eingabe und Antwort erschienen mir ...
... sehr lang [5] ... sehr kurz

1	2	3	4	5	6
18	46	25	3	0	0

keine Wertung: 0

Frage 25

Mit Hilfe der Spracheingabe komme ich schneller ans Ziel
trifft nicht zu [6] trifft zu

1	2	3	4	5	6
23	22	11	15	12	6

keine Wertung: 3

Frage 26

Die Kombination von Stift- und Spracheingabe finde ich sinnvoll.
trifft nicht zu [6] trifft zu

1	2	3	4	5	6
1	4	2	8	28	48

keine Wertung: 1

Frage 27

Im Gespräch mit SmartWeb fühlte ich mich ...
... unwohl [6] ... wohl

1	2	3	4	5	6
0	8	13	26	29	15

keine Wertung: 1

Frage 28

Ich bin von der Leistung von SmartWeb
... enttäuscht [6] ... beeindruckt

1	2	3	4	5	6
15	25	20	20	12	0

```
keine Wertung: 0
================================================================================

================================================================================
Frage 29
--------------------------------------------------------------------------------
Der Umgang mit SmartWeb hat ...
mich gelangweilt [6] mir Spaß gemacht
-----------------------------
| 1 | 2 | 3  | 4  | 5  | 6  |
|---|---|----|----|----|----|
| 4 | 8 | 12 | 20 | 23 | 25 |
-----------------------------
keine Wertung: 0
================================================================================
```

8.5 Task Level Documents

Aufgabenstellungen zur 2. Sitzung – Städteausflug
¨

In der heutigen Sitzung sind 4 Aufgabenstellungen mit den Themen Wetter, Navigation, Sehenswürdigkeiten und Ubernachtung zu bearbeiten. Vor der nächsten Eingabe sollte der Sprach-Output abgewartet werden.

Rahmengeschichte: Sie planen ab morgen einen Kurztripp nach Berlin. Benutzen Sie SmartWeb, um sich vor der Reise Auskünfte zur Reiseplanung einzuholen. Falls Sie noch Fragen haben, wenden Sie ich bitte an den Versuchsleiter.

Verbindungstyp WLAN
Umgebung Büro
Headset nein
Mikro ptt
indoor/outdoor indoor

Aufgabe I: Wetter

1. Erkundigen Sie sich nach dem morgigen Wetter in Berlin.

2. Und nun indem Sie SmartWeb den morgigen Wochentag nennen.

3. Und nun nach dem Wetter in zwei Tagen.

Aufgabe II: Navigation

1. Erkundigen Sie sich wie Sie nach Berlin kommen.

2. Lassen Sie sich ein Karte von Berlin zeigen.

3. Arbeiten Sie mit der Karte. Probieren Sie z.B. die Zoom-Funktion aus.

Aufgabe III: Sehenswürdigkeiten in Berlin: Brandenburger Tor, Alexanderplatz, Fernsehturm, Bahnhof Zoo, Zoologischer Garten.

1. Lassen Sie sich Informationen zu einer der oben genannten Sehenswürdigkeiten geben.

2. Lassen Sie sich ein Bild von dieser Sehenswürdigkeit zeigen.

3. Und nun fragen Sie nach einer anderen bekannten Sehenswürdigkeit.

Aufgabe IV: Übernachtung

1. Fragen Sie nach Hotels in Berlin

2. Wenn Smartweb Ihnen eine Karte anzeigt, klicken Sie auf das Textfeld Antworten und lassen sich eine Liste der Hotels anzeigen. Scrollen Sie die Liste durch.

3. Suchen Sie sich eines der Hotels aus und fragen Sie SmartWeb nach einem Bild des Hotels.

Aufgabenstellungen zur 3. Sitzung - Fußball

In der heutigen Sitzung steht das Thema Fußball im Mittelpunkt. Die einzelnen Aufgabenstellungen beziehen sich dabei auf die Themen Fußballnationalmannschaft, Weltmeisterschaften und Fußballgeschichte. Zur Eingabe Ihrer Anfragen können Sie die Möglichkeit der Spracheingabe oder der Stifteingabe verwenden. Vor der nächsten Eingabe sollte die Sprachausgabe des Systems abgewartet werden. Versuchen Sie immer eine Aufgabe komplett abzuschließen.

Rahmengeschichte: Sie interessieren sich für verschiedene Fakten des Themengebietes Fußball und wollen dabei insbesondere Informationen zur vergangenen Weltmeisterschaft hinsichtlich Austragungsorte, Anfahrtswege, qualifizierte Mannschaften, Siegermannschaften etc. einholen. Falls Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Verbindungstyp UMTS
Umgebung Straße
Headset ja
Mikro ptt
indoor/outdoor outdoor

Aufgabe I: Fußballnationalmannschaft

1. Erkundigen Sie sich bei SmartWeb nach den deutschen Fußballnationalspielern

2. Lassen Sie sich Bilder von bekannten Spielern zeigen.

3. Fragen Sie nun nach Toren dieses Spielers.

Aufgabe II: Weltmeisterschaft

1. Sie interessieren sich dafür, wann und wo die letzte Fußball-WM stattgefunden hat.

2. Erkundigen Sie sich nach dem Austragungsort des Endspiels.

3. Lassen Sie sich eine Karte des Spielortes zeigen.

4. Arbeiten Sie mit der Karte. Probieren Sie z.B. die Zoom-Funktion aus.

Aufgabe III: Historie Austragungsorte der vergangenen Weltmeisterschaften, Titelträger, Torschützenkönige, Mannschaften

1. Lassen Sie sich Informationen zu einem der oben genannten Schlagworte geben.

2. Lassen Sie sich ein Bild des letzten Weltmeisters anzeigen.

3. Fragen Sie nun nach Interviews mit einem bekannten Spieler.

4. Sie interessieren sich für die Maskottchen der Weltmeisterschaften.

Aufgabe IV: Anfahrt und Übernachtung
Sie wollen zu einem wichtigen Fußballspiel nach Hamburg fahren.

1. Fragen Sie nach günstigen Reiserouten nach Hamburg.

2. Suchen Sie dort nach Hotels.

Aufgabenstellungen zur 4. Sitzung - Unterhaltung

In der heutigen Sitzung stehen Unterhaltung und Verpflegung im Mittelpunkt. Die vier Aufgabenstellungen beziehen sich dabei auf die Bereiche Kino, Kultur, Gastronomie und ärztliche Versorgung. Zur Eingabe Ihrer Anfragen können Sie die Möglichkeit der Spracheingabe oder der Stifteingabe verwenden. Vor der nächten Eingabe sollte die Sprachausgabe des Systems abgewartet werden.

Rahmengeschichte: Sie haben vor, sich heute einen schönen Abend zu machen. Da Sie weder über Kultur noch über Restaurants und Bars in München auf dem Laufenden sind, wollen Sie verschiedene Informationen mit Hilfe von Smartweb einholen. Planen Sie mit SmartWeb Ihren Abend. Falls Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Verbindungstyp UMTS
Umgebung Cafe
Headset ja
Mikro ptt
indoor/outdoor indoor

Aufgabe I: Wahl des Kinos

1. Erkundigen Sie sich bei SmartWeb nach Kinos in München.

2. Lassen Sie sich einen Stadtplan mit den Kinos anzeigen.

3. Arbeiten Sie mit dieser Karte.

Aufgabe II: Kultur- und Unterhaltungsprogramm

1. Lassen Sie sich eine Liste der aktuellen Kinofilme zeigen.

2. Fragen Sie SmartWeb nach aktuellen Veranstaltungen bzw. nach einem Veranstaltungskalender.

Aufgabe III: Gastronomie

1. Fragen Sie SmartWeb nach Restaurants in München.

2. Grenzen Sie ihre Frage auf Italiener oder Steakhäuser oder griechische Restaurants ein.

3. Nach dem Essen möchten Sie spazieren gehen. Fragen Sie SmartWeb nach dem Englischen Garten oder dem Olympiapark.

4. Um ihr Auto abstellen zu können, fragen Sie SmartWeb nach einem Parkhaus.

Aufgabe IV: Ärzte, Apotheken
Leider haben Sie sich nach einem deftigen Essen den Magen verdorben. Um den Abend zu retten, fragen Sie Smartweb nach Hilfe.

1. Erkundigen Sie sie nach Ärzten in München.

2. Und nun nach Apotheken.

Aufgabenstellungen zur 5. Sitzung - Planung einer Deutschlandrundreise

Ihre Aufgabe in der heutigen Sitzung ist, sich Information über Deutschland für eine geplante Rundreise einzuholen. Die Aufgaben sind wieder in vier Themenbereiche unterteilt: Geographie, Politik, Geschichte und Deutschlands berühmte Persönlichkeiten. Zur Eingabe Ihrer Anfragen können Sie die Möglichkeit der Spracheingabe oder der Stifteingabe verwenden. Vor der nächsten Eingabe sollte die Sprachausgabe des Systems abgewartet werden.

Rahmengeschichte: Ihr Wissen über Land und Leute ist sehr gering. Bevor Sie Ihre Rundreise in die Tat umsetzen, möchten Sie Ihre Wissenslücken auffüllen und recherchieren mit Hilfe von SmartWeb nach wissenswerten Fakten über Deutschland. Falls Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Verbindungstyp WLAN
Umgebung Büro
Headset nein
Mikro ptt
indoor/outdoor indoor

Aufgabe I: Geographie

1. Erkundigen Sie sich nach Deutschlands Bundesländern.
2. Fragen Sie nach dem Namen der Hauptstadt.
3. Fragen Sie nach der Einwohnerzahl der Hauptstadt.

Aufgabe II: Geschichte

1. Erkundigen Sie sich über die Gründung der Bundesrepublik Deutschuland.

2. Und nun, wann die DDR gegründet wurde.

Aufgabe III: Politik

1. Fragen Sie nach Parteien in Deutschland.

2. Informieren Sie sich, wer Bundesklanzler/in von Deutschland ist.

3. Fragen Sie SmartWeb, wann diese/r geboren ist oder Bundeskanler/in wurde oder nach einem Bild der Person.

4. Stellen Sie eine Frage über den Reichstag.

Aufgabe IV: berühmte Persönlichkeiten Franz Beckenbauer, Michael Ballack, Goethe

1. Stellen Sie eine beliebige Frage über die Person ihrer Wahl.

2. Lassen Sie sich ein Bild von ihr/ ihm zeigen.

3. Fragen Sie konkret nach einem Werk bzw. Tore, oder Alter / Tod/ Geburt.

Aufgabenstellungen zur 6. Sitzung - Start der Deutschlandrundreise

Thema der heutigen Sitzung ist die Umsetzung Ihrer in der letzten Sitzung geplanten Deutschlandreise. Sie bearbeiten wieder vier Aufgaben zu folgenden Themen: Verkehrsinformation, Wetter, Sehenswürdigkeiten und Freizeit. Vor der nächten Eingabe sollte die Sprachausgabe des Systems abgewartet werden.

Rahmengeschichte: Nachdem Sie sich über Deutschland informiert haben, starten Sie heute Ihre Rundreise. Ausgangspunkt ist München, von dort aus möchten sie weiter in Richtung Norden reisen. Falls Sie noch Fragen haben, wenden Sie sich bitte an den Versuchsleiter.

Verbindungstyp WLAN
Umgebung Büro
Headset ja
Mikro offen
indoor/outdoor indoor

Aufgabe I: Verkehrsinformation

Summative Evaluation of the SmartWeb Prototype 1.0

1. Fragen Sie nach einer Route von München nach Stuttgart.

2. Erkundigen Sie sich nach der Verkehrslage bzw. nach Staus auf der Strecke München - Stuttgart.

3. Und eine Route von Stuttgart nach Berlin.

Aufgabe II: Wetter

1. Sie können sich nicht entscheiden, welche Stadt Ihr nachstes Reiseziel sein soll. Darum fragen Sie SmartWeb nach dem Wetter in Hamburg.

2. Und nun nach dem Wetter in Berlin.

3. Und nun nach dem Wetter in einer anderen nördlich von München gelegenen Großstadt.

Aufgabe III: Sehenswürdigkeiten

1. Sie haben Sich für Berlin entschieden. Erkundigen Sie sich, wo sich der Potsdamer Platz oder der Palast der Republik befindet.

2. Lassen Sie sich Bilder von einer der oben genannten Sehenswürdigkeiten zeigen.

3. Fragen Sie SmartWeb, wann der Palast der Republik gebaut wurde oder wer der Architekt ist.

Aufgabe IV: Freizeit

1. Das Bargeld wird knapp. Nutzen Sie SmartWeb, um eine Bank in Berlin zu finden.

2. Sightseeing macht Hunger. Sie möchten mit Hilfe von SmartWeb ein indisches Restaurant in Berlin finden.

3. Und nun Kinos in der Stadt.

8.6 Tested Conditions

The following table lists all performed evaluation session together with their most prominent conditions. The columns from left to right are: task level, session ID, connection type (WLAN/UMTS), environment (indoor/outdoor), headset type (none, MDA_Pro), ASR activation type per buttom push (ptt) or open microphone (active) and test location ('Büro' = office, Cafe, 'Straße' = street).

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
0	e097	WLAN	indoor	MDA_Pro	ptt	Büro
0	e098	WLAN	indoor	none	ptt	Büro
0	e099	WLAN	indoor	none	ptt	Büro
0	e100	WLAN	indoor	none	ptt	Büro
0	e104	WLAN	indoor	none	ptt	Büro
0	e105	WLAN	indoor	none	ptt	Büro
0	e106	WLAN	indoor	none	ptt	Büro
0	e107	WLAN	indoor	none	ptt	Büro
0	e110	WLAN	indoor	none	ptt	Büro
0	e111	WLAN	indoor	none	ptt	Büro
0	e112	WLAN	indoor	none	ptt	Büro
0	e116	WLAN	indoor	none	ptt	Büro
0	e117	WLAN	indoor	none	ptt	Büro
0	e128	WLAN	indoor	none	ptt	Büro
0	e134	WLAN	indoor	none	ptt	Büro
0	e157	WLAN	indoor	none	ptt	Büro

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
1	e101	WLAN	indoor	none	active	Büro
1	e102	WLAN	indoor	MDA_Pro	ptt	Büro
1	e103	WLAN	indoor	MDA_Pro	ptt	Büro
1	e109	WLAN	indoor	none	ptt	Büro
1	e113	WLAN	indoor	none	ptt	Büro
1	e114	WLAN	indoor	none	ptt	Büro
1	e115	WLAN	indoor	none	ptt	Büro
1	e119	WLAN	indoor	none	ptt	Büro
1	e121	WLAN	indoor	none	ptt	Büro
1	e124	WLAN	indoor	none	ptt	Büro
1	e126	WLAN	indoor	none	ptt	Büro
1	e130	WLAN	indoor	none	ptt	Büro
1	e132	WLAN	indoor	none	ptt	Büro
1	e135	WLAN	indoor	none	ptt	Büro
1	e138	WLAN	indoor	none	ptt	Büro
1	e159	WLAN	indoor	none	ptt	Büro

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
2	e118	UMTS	outdoor	none	ptt	Straße
2	e120	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e123	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e125	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e127	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e131	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e133	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e136	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e137	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e139	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e140	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e141	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e142	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e145	UMTS	outdoor	MDA_Pro	ptt	Straße
2	e165	WLAN	indoor	MDA_Pro	ptt	Cafe

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
3	e122	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e129	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e143	UMTS	indoor	MDA_Pro	ptt	Cafe
3	e144	UMTS	indoor	MDA_Pro	ptt	Cafe
3	e150	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e151	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e153	UMTS	indoor	MDA_Pro	ptt	Cafe
3	e154	UMTS	indoor	MDA_Pro	ptt	Cafe
3	e155	UMTS	indoor	MDA_Pro	ptt	Cafe
3	e158	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e161	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e163	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e164	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e168	UMTS	outdoor	MDA_Pro	ptt	Cafe
3	e176	UMTS	indoor	MDA_Pro	ptt	Cafe

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
4	e146	WLAN	indoor	none	ptt	Büro
4	e147	WLAN	indoor	none	ptt	Büro
4	e148	WLAN	indoor	none	ptt	Büro
4	e149	WLAN	indoor	none	ptt	Büro
4	e152	WLAN	indoor	none	ptt	Büro
4	e156	WLAN	outdoor	none	ptt	Büro
4	e160	WLAN	indoor	none	ptt	Büro
4	e166	WLAN	indoor	none	ptt	Büro
4	e167	WLAN	indoor	none	ptt	Büro
4	e169	WLAN	indoor	none	ptt	Büro
4	e171	WLAN	indoor	none	ptt	Büro
4	e177	WLAN	indoor	none	ptt	Büro
4	e179	WLAN	indoor	none	ptt	Büro
4	e181	WLAN	indoor	none	ptt	Büro
4	e185	WLAN	indoor	none	ptt	Büro
4	e187	WLAN	indoor	none	ptt	Büro

e_task_number	s_session_name	s_connection_type	s_situation	s_headsettype	s_asr_state	s_directions_location
5	e162	WLAN	indoor	none	active	Büro
5	e170	WLAN	indoor	none	active	Büro
5	e173	WLAN	indoor	none	active	Büro
5	e178	WLAN	indoor	MDA_Pro	active	Büro
5	e180	WLAN	indoor	none	active	Büro
5	e182	WLAN	indoor	MDA_Pro	active	Büro
5	e183	WLAN	indoor	MDA_Pro	active	Büro
5	e184	WLAN	indoor	MDA_Pro	active	Büro
5	e186	WLAN	indoor	none	active	Büro
5	e188	WLAN	indoor	none	active	Büro
5	e190	WLAN	indoor	none	active	Büro
5	e191	WLAN	indoor	MDA_Pro	active	Büro
5	e193	WLAN	indoor	MDA_Pro	active	Büro
5	e194	WLAN	indoor	MDA_Pro	active	Büro

8.7 Questionnaire A – Results

In the following you will find the raw results from the initial questionnaire A answered by each test subject prior to the first test. Answers with 0 hits are not shown.

uea_experience_internet	count
ein paar Mal pro Woche	1
etwa einmal pro Tag	6
mehrmals am Tag	9

uea_experience_diasys	count
Ich entwickele selber Dialogsysteme	1
Ich verwende nur selten ein Dialogsystem	9
Ich verwende sie jede Woche	2
Ich verwende sie nicht mehr als einmal im Monat	4

uea_pre_opinion_diasys	count
1 ... trotzdem lieber mit einem Menschen	6
2	4
3	4
4	2

uea_pre_payfor_diasys	count
bereit, 10 Cent pro Minute zu zahlen	8
bereit, 50 Cent pro Minute zu zahlen	1
nicht bereit, etwas zu zahlen	7

uea_pre_opinion_service	count

```
-------------------------+-------
 Stimme ich voll zu      |     5
 Tendenziell falsch      |     1
 Tendenziell richtig     |     6
 Weitgehend richtig      |     4

 uea_pre_opinion_human | count
-----------------------+-------
 Stimme ich voll zu    |     8
 Tendenziell falsch    |     2
 Tendenziell richtig   |     4
 Weitgehend richtig    |     2

 uea_pre_opinion_personalassi | count
------------------------------+-------
 Tendenziell falsch           |     6
 Tendenziell richtig          |     2
 Weitgehend falsch            |     1
 Weitgehend richtig           |     7

   uea_pre_opinion_help   | count
--------------------------+-------
 Finde ich extrem wichtig |     9
 Ziemlich wichtig         |     7

 uea_synthesis_simple | count
----------------------+-------
 einige Male          |     1
 oft                  |     4
 sehr oft             |    11

 uea_synthesis_sms | count
-------------------+-------
 einige Male       |     6
 nie               |     2
 oft               |     1
 sehr oft          |     1
 selten            |     6

 uea_synthesis_nav_simple | count
--------------------------+-------
 einige Male              |     4
 nie                      |     3
 oft                      |     2
 sehr oft                 |     3
 selten                   |     4

 uea_synthesis_nav_complex | count
---------------------------+-------
 einige Male               |     2
 nie                       |     3
 oft                       |     2
 sehr oft                  |     5
 selten                    |     4
```

8.8 Questionnaire C – Results

In the following you will find the raw results from the initial questionnaire C answered by each test subject right after the last test. Answers with 0 hits are not shown.

uec_post_payfor_diasys	count
bereit, 10 Cent pro Minute zu zahlen	11
bereit, 25 Cent pro Minute zu zahlen	2
nicht bereit, etwas zu zahlen	3

uec_post_opinion_service	count
Stimme ich voll zu	6
Tendenziell falsch	3
Tendenziell richtig	2
Weitgehend richtig	5

uec_post_opinion_human	count
Stimme ich voll zu	7
Tendenziell falsch	1
Tendenziell richtig	3
Weitgehend richtig	5

uec_post_opinion_personalassi	count
Stimme ich voll zu	2
Stimme ich überhaupt nicht zu	1
Tendenziell falsch	4
Tendenziell richtig	3
Weitgehend falsch	1
Weitgehend richtig	5

uec_post_opinion_help	count
Stimme ich voll zu	13
Tendenziell richtig	1
Weitgehend richtig	2

uec_voice_fit_to_system	count
Gut	5
Ordentlich	7
Schlecht	3
Sehr schlecht	1

uec_voice_quality	count
Gut	3
Ordentlich	7
Schlecht	4
Sehr gut	1
Sehr schlecht	1

uec_voice_pleasentness	count
Angenehm	2
Neutral	7
Unangenehm	7

uec_voice_naturalness	count
Neutral	7
Sehr unnatürlich	2
Unnatürlich	7

```
 uec_voice_applicability | count
-------------------------+-------
 Ja                      |     9
 Nein                    |     7

 uec_additional_questions | count
--------------------------+-------
 Stimme ich voll zu       |    15
 Tendenziell richtig      |     1

    uec_defined_formulation    | count
-------------------------------+-------
 Stimme ich voll zu            |     4
 Stimme ich überhaupt nicht zu |     2
 Tendenziell falsch            |     1
 Tendenziell richtig           |     1
 Weitgehend falsch             |     4
 Weitgehend richtig            |     4

       uec_know_infotype       | count
-------------------------------+-------
 Stimme ich voll zu            |     7
 Stimme ich überhaupt nicht zu |     1
 Tendenziell falsch            |     1
 Tendenziell richtig           |     2
 Weitgehend richtig            |     5

   uec_acknowledge   | count
---------------------+-------
 Stimme ich voll zu  |    10
 Tendenziell richtig |     2
 Weitgehend richtig  |     4

   uec_processing   | count
--------------------+-------
 Stimme ich voll zu |    16
(1 row)

  uec_modify_input  | count
--------------------+-------
 Stimme ich voll zu |    16
(1 row)
```