D6.1: Proposed Methods for Multimodal Experiments

Tilman Becker, Oliver Lemon, and Steve Young (editors),
Nate Blaylock, Ivana Kruiff-Korbayova, Kalliroi Georgila,
James Henderson

Distribution: Public

TALK
Talk and Look: Tools for Ambient Linguistic Knowledge
IST-507802 Deliverable 6.1

November 15th, 2004

Project funded by the European Community
under the Sixth Framework Programme for
Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref. no.	IST-507802
Project acronym	TALK
Project full title	Talk and Look: Tools for Ambient Linguistic Knowledge
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 January 2004 / 36 Months

Security	Public
Contractual date of delivery	M9 = September 2004
Actual date of delivery	November 15th, 2004
Deliverable number	6.1
Deliverable title	D6.1: Proposed Methods for Multimodal Experiments
Type	Report
Status & version	Final 1.0
Number of pages	24 (excluding front matter)
Contributing WP	6
WP/Task responsible	DFKI
Other contributors	All partners
Author(s)	Tilman Becker, Oliver Lemon, and Steve Young (editors), Nate Blaylock, Ivana Kruiff-Korbayova, Kalliroi Georgila, James Henderson
EC Project Officer	Kimmo Rossi
Keywords	experimental methods, mulitmodal experiments, evaluation, data collection

The partners in TALK are:

Saarland University	USAAR
University of Edinburgh HCRC	UEDIN
University of Gothenburg	UGOT
University of Cambridge	UCAM
University of Seville	USE
Deutches Forschungszentrum fur Künstliche Intelligenz	DFKI
Linguamatics	LING
BMW Forschung und Technik GmbH	BMW
Robert Bosch GmbH	BOSCH

For copies of reports, updates on project activities and other TALK-related information, contact:

The TALK Project Co-ordinator
Prof. Manfred Pinkal
Computerlinguistik
Fachrichtung 4.7 Allgemeine Linguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
pinkal@coli.uni-sb.de
Phone +49 (681) 302-4343 - Fax +49 (681) 302-4351

Copies of reports and other material can also be accessed via the project's administration homepage, `http://www.talk-project.org`

Contents

Executive summary

This report details work done in the early phases of the TALK project to establish an appropriate evaluation framework for multimodal dialogue systems and their components, with respect to the TALK project research goals.

Chapter 1

Introduction

The TALK project is concerned with the design of multimodal dialogue systems where speech is a primary input modality, with gestures (mouse, joystick, pen) as and additional input modality, and a mixture of speech and graphics is used for dialogue system output. This report details work done in the early phases of the project to establish an appropriate evaluation framework for such systems and their components, with respect to the TALK project research goals.

The block diagram of a typical multimodal dialogue system is shown in Figure 1.1. As can be seen, the system consists of several components, each of which is relatively complex. During the lifetime of the project many of these components will evolve and in general a complete system will only be available towards the end of the project. Although testing at the component level can provide valuable data before then, some key questions can only be answered by data obtained from a complete system.

To deal with these problems, the TALK project has adopted the following broad strategies:

- identify the key issues at an early stage
- test at the component level where possible
- collect representative data from "Wizard of Oz" (WoZ) systems at an early stage in the project
- build preliminary evaluation systems at the mid-point in the project
- complete full-system testing in the final phase of the project.

Although there will be a number of different laboratory systems built, all of the systems will focus on either in-car or in-house scenarios. In-car scenarios will include route planning, tourist information, personal information management and MP3 audio. The in-house scenario will include control of domestic systems and entertainment including MP3 audio.

This document reports work done in the project to address the first three of these strategies, and comments briefly on the latter two. The organisation of the document is as follows.

Firstly, chapter 2 summarizes the key research questions and evaluation methodologies which will be addressed at the component level. Much of the experimental work to be done at this level is already underway, and the outcomes will be reported in the appropriate work packages (indicated using square brackets).

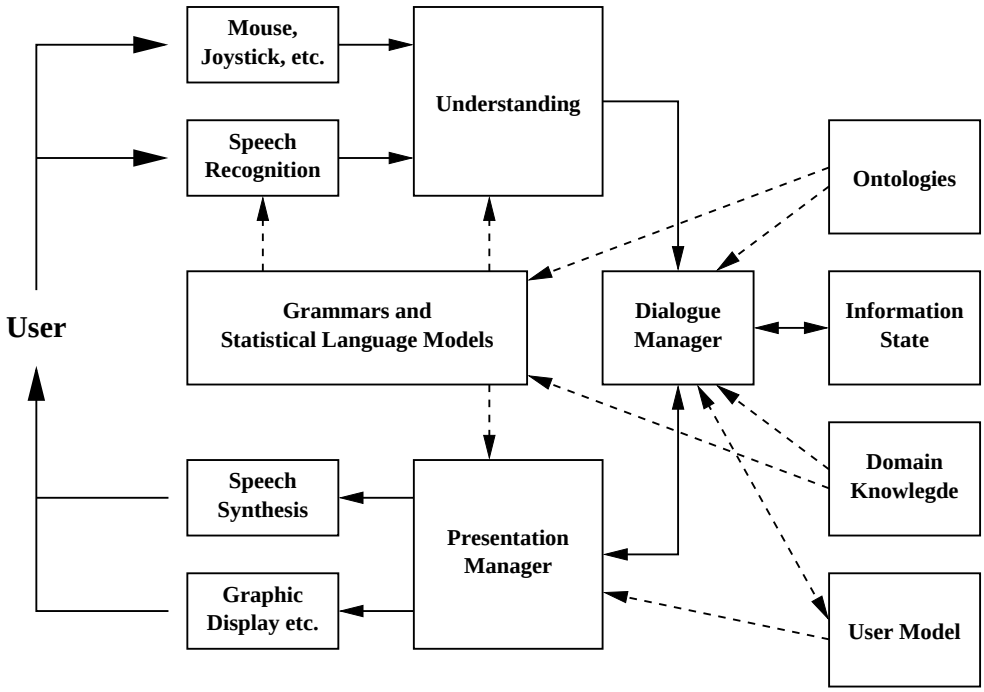

Figure 1.1: Main Components in a Multi-modal Dialogue System

Secondly, chapter 3 describes how WoZ data collection is performed and the uses that such data can be put to. Again, although much has already been achieved in this area, the work is ongoing and will continue throughout the project.

Thirdly, chapter 4 briefly describes the main systems that will be built and used for final evaluation and testing.

Finally, chapter 5 presents a schedule for the proposed experiments.

Chapter 2

Component Level Evaluation

Evaluations of single components typically are contrastive evaluations between two versions of a component that generally vary a single parameter. In most cases this is a binary distinction, e.g., a statistical vs. a grammar-based language model, in some cases the experiments are expected to determine an optimal value for a numerical parameter, e.g., the number of examples that are presented from large sets of results to a user query. The following sections present the experiments for components that are developed in workpackages 1 to 4.

2.1 Grammars and Language Models

Grammars and statistical language models (SLMs) are used for three specific functions:

- priming the speech recogniser with a set of allowable grammatical forms
- parsing the output of the speech recogniser to form an interpretation
- informing the generation of speech/text output.

The second of these is dealt with in section 2.2 on understanding and the last of these is dealt with below in section 2.3 on presentation.

The use of grammars and statistical language models for priming a recogniser introduces a number of difficult issues into dialogue system design. An explicit grammar can provide precise and accurate recognition and interpretation within domain, but is fragile when the user strays out of domain. An SLM is significantly more robust but is less precise and requires large amounts of training data. These issues are further complicated by the incorporation of multi-modal input streams and the desire to generate grammars automatically via the use of abstract grammar formalisms and ontologies.

Within the TALK project, the following specific aspects of the above will be investigated:

- the integration of grammars and SLMs and the trade-offs on performance [WP1.3]
- the integration of multimodal inputs (e.g. click on a map) both explicitly within abstract grammars and implicitly by incorporating into SLMs. [WP1.4]

- the interaction of grammar and modality coordination with existing knowledge representations and domain sources [WP1.5, WP2.1]
- automatic configuration of grammars using ontologies [WP2.1] and "voice programming" of devices [WP2.3]

Testing grammars and statistical language models at the component level is typically done by measuring one or more of word error rate (WER), semantic concept error rate, and semantic concept F-measure where the latter combines the ability to retrieve (R) semantic concepts from the input and the precision (P) with which this retrieval is done ($F = \frac{2PR}{P+R}$). These measures will be conducted throughout the project to evaluate the various issues described above. Where multimodal input is involved, an effective approach is to compare understanding performance (whether at the word or semantic concept level) with and without the additional modes enabled.

2.2 Understanding

Semantic interpretation of user inputs typically depends on grammars for parsing and domain constraints for resolving ambiguities. These processes can also be used to select from alternative input hypotheses thus indirectly improving the effective recognition rate. Multimodal data streams can be integrated before parsing and interpreted by a single unified process or they can be interpreted separately and the results integrated in a post-processing phase. Also, as in the previous section, grammars can be either hand-crafted or generated automatically from abstract specifications, "plug and play" library modules and ontologies. The trade-off between these is an important research question for practical systems.
Within the TALK project, the following specific aspects of the above will be investigated and evaluated:

- comparison of hand-crafted vs statistical parsing models [WP1.4]
- automatic configuration and use of ontologies including comparison with hand-crafted systems [WP2.1]
- using semantic/contextual appropriateness to re-rank speech recognition N-best output [WP2.1]
- user-driven extension of semantic processing rules using "voice programming" [WP2.3].

Testing system understanding at the component level can share the same F-measure metrics used in section 2.1 and the re-ranking experiments can measure effective word error rate.

2.3 Presentation

Presentation involves a variety of issues including choice of modalities, information summarisation, information structuring, planning, layout and natural language generation. At the user level, the following research questions will be of particular interest:

- How can complex information (e.g., lists of options, trip information) be distributed effectively on the available modalities?

- How does the multimodal presentation of information affect the user's workload and performance in a multi-task setting such as in the in-car domain?
- Does the user change modalities during the dialogue and does the form of output influence this?

These issues will be addressed for the in-car domain by conducting WoZ experiments using a driving simulator (i.e., the Lane Changing Task) and also by trials using the baseline evaluation system (see chapter 4 below). Measurements taken will include reaction time, control errors, task completion rate, and dialogue effectiveness.

Planning multimodal turns for dialogue poses additional decision problems to those of a "pure text" generation system [CDE+99]. In order to make appropriate decisions on the modality type(s) and on the information distribution over the chosen modalities and the amount of redundancy, we are interested in different facettes of empirical experiments. The overall question is to see whether—and if so, how—multimodal output increases the efficiency of communication in different contexts. We would like to experiment with different, systematically varied combinations and realizations of multimodal output in order to get an insight into how to present information in diverse situations in an optimal way. However, the multimodal combinations and changes should appear as natural as possible.

Turning now to the specific issues relating to the presentation system component, the following aspects will be investigated and evaluated:

- representation of modality coordination in the information state [WP3.1]
- generating referring expressions in intra-modal, inter-modal and multi-lingual contexts [WP3.2]
- efficient presentation of information, content selection and use of redundancy [WP3.2]
- flexible utterance realization. Is it useful to change the dialogue output to match the context? Is priming and alignment an adequate way of constraining the realization of the utterance? [KKERK03] [WP3.2]
- how should modes be synchronized? E.g., should speech come before, during or after graphical display? How sensitive does the synchronization need to be? [WP3.2]
- displaying the Dialogue State. Is it helpful to give feedback on the Dialogue State to the user and if so how should this be done? [WP3.2]
- attention distribution over modalities. Does dividing the user's attention between modalities reduce efficiency of information processing or communication [Ovi99]? [WP3.2]
- attentivity for communication as secondary activity. How is attention distribution affected when there is a disruptive primary task such as driving? [WP3.2]
- user information selection strategies. How should dialogue be designed to allow a user to select different pieces of information or to narrow down a selection? [WP3.2]
- adaptive use of graphical elements. What is the best way to lay out larger sets of data, e.g., in tables? What level of detail should be used? [WP3.2, WP3.3].

Unlike the input components whose performance can be measured using objective metrics, evaluation of the presentation component will rely mostly on subjective testing and system-level evaluation metrics such as, e.g., task-completion rate. This can be done in one of two ways:

- general system evaluation comparing the prototypes with a competing or a baseline system.
- Wizard-of-Oz (WoZ) type experiments.

Component level evaluations can be conducted by presenting alternatives for system output and request judgments on appropriateness, either on an absolute scale or at least as relative judgments, comparing alternatives. Where WoZ experiments are used, important information will be available from both the Wizard and the User. The former will need to make presentational choices and these will inform the work of WP3.2 which is concerned with developing modality-specific resources. Section 3.1.2 describes an initial experiment in this paradigm. Evaluation of user performance and preferences will inform WP3.1 where the emphasis is on using the information state to plan the presentation of information.

2.3.1 Determining Presentation Planning Rules and Parameters

As a concrete example, this section sketches the experimental setup for a particular research question in multimodal presentation planning. When a database query is underspecified (e.g., titles from a well-known artist in the FreeDB music database), the number of results can be quite large. While on a desktop the solution might be large virtual window with a scrollbar, this is not an option in speech based dialogue and not an option with restricted graphical displays as in a car (or any other mobile device) and not an option in a situation where the user attentiveness is lowered by a primary task (e.g., driving a car).

The initial WoZ experiments described in section 3.1.2 have shown that human subjects use a number of different strategies: summarize the result (e.g., give the number of database query results), present partial results (e.g., present the first few results), and offer refinements. To help determine an optimal presentation strategy, the experiment sketched in the following has two goals:

1. to determine the best presentation strategy
2. to determine, under the assumption that a partial presentation of results is generally desirable, the best number of results to be presented (parameter optimization).

These questions are addressed by single turns in which the user is guided to make an underspecified request and the system or wizard chooses one of, e.g., the following three strategies:

- (i) top-down (or summary first): *There are 100 titles. The first 3 are: ... Select one or ask for more.*
- (ii) bottom-up (or data first): *I've found: ... There are 97 more. Select one or ask for more.*
- (iii) offer refinement: *There are too many titles to list. Do you prefer older or newer songs?*

All three strategies have many potential variants (e.g. (ii) could offer contrastive descriptions of items as in [MFLW04]) and must be chosen with care. Within the first two strategies, the number of titles presented will be varied to determine the optimal value for this parameter. Since the length of titles can vary considerably, the total time to speak the list of titles (or as an approximation, the number of characters) must be taken into account besides the bare number of titles.

This core dialogue turn that is of experimental interest must be embedded in a larger task, e.g., to construct a playlist with one song from each of five given artists. Experimental measures that can be used to determine the appropriateness of each strategy include component-level data, i.e., wizard choices and

user judgments as well as system-level data, e.g. on task completion rate, time (and number of turns) to complete the task, user satisfaction (as queried afterwards), number of error turns, etc.

Further questions to be addressed in these experiments include multimodal issues such as modality assignments and preferences, redundancy and multimodal user models.

Follow-up experiments can then be conducted, e.g., varying environmental factors such as the workload from the primary task or testing the impact of a user model (e.g., sorting the list of results and first presenting titles which are known by the user).

2.4 Dialogue Management and Adaptivity

The dialogue manager is the core control component of a multimodal dialogue system. Based on the current information state and each user input, a new output action must be determined and the information state updated. Much of the focus in TALK (WP4) is on using machine learning techniques to optimise dialogue managers during development and to further adapt them on-line as they are exposed to real user interactions. These techniques work by modelling the dialogue as a Markov Decision Process (MDP), assigning rewards (both positive and negative) to various dialogue states and then finding the dialogue policy which leads to the maximum expected reward starting from any dialogue state [You00].

Within the TALK project, the following specific aspects of the above will be investigated:

- which state variables are particularly relevant to determining low level dialogue confirmation and information presentation strategies? [WP4.2]
- what is the effect of different reward signals on resulting dialogue policy? [WP4.2]
- how can the large ISU state space be mapped effectively into a tractable subset for reinforcement learning (RL)? (e.g., using Bayes Nets, linear function approximation, etc) [WP4.3]
- how can the high computational cost of learning in large state spaces be reduced? (e.g., via function approximation, algebraic decision diagrams, etc.) [WP4.3]
- do partially observable MDPs (POMDPS) offer any advantages compared to MDPs? [WP4.4]
- how can the tractability issues inherent in POMDPs be avoided? [WP4.4].

Some testing of the dialogue manager can be performed at the component level by simulating the remaining components at the intention level (e.g., as in [SY01]). In such cases, appropriate metrics are task completion rates and times, number of turns to achieve completion, number of times asked for help, user satisfaction etc. In the context of reinforcement learning (RL), dialogues can be simulated and rewards measured for different conditions. The performance of optimised dialogues can then be contrasted with hand-coded strategies.

The dialogue manager can also be tested by embedding it in a prototype system and running user trials. This will be the main evaluation approach and the primary evaluation will compare the optimised system with a hand-crafted baseline. UEDIN plans to conduct such experiments midway through the project (i.e., in mid-2005). Similar metrics to the above can still be used, but in this case subjective measures can be gathered from user questionnaires.

2.5 Resources for Component Level Evaluation

For the input components, the main requirements are test input data in the form of user inputs (spoken utterances, gestures, etc) collected and transcribed. In TALK both existing and new data sources are being used. Existing data includes the classic ATIS [DBB+94] and Communicator corpora [WAB+01], and data collected by the partners in previous projects. New data includes that obtained in WoZ experiments (see chapter 3) for both the MP3 audio play task and the in-car tourist information task.

In the case of dialogue manager optimisation, dialogue corpora annotated at the turn level with dialogue acts and information states can be used for machine learning experiments. Within TALK such data has been obtained from the WoZ experiments and by exploiting existing corpora. For the latter, a particularly rich source is the DARPA Communicator archive mentioned above. However to make this useful for TALK it is necessary to annotate the dialogues with a sequence of information states which can then be used to feed machine learning algorithms. The Communicator data is a large corpus of dialogues regarding flight booking tasks as well as car rentals and hotel reservations. It therefore shares much in common with the in-car scenario and the lessons learnt from processing this data will have direct relevance to the project goals.

2.5.1 Annotation of the Communicator data

The whole process of annotating the Communicator data with information states has been implemented using DIPPER [BKLO03] and OAA [CM01]. Two OAA agents have been developed. The first one is used for reading the Communicator XML file, which contains all the information about dialogues, turns, utterances, transcriptions, etc. Each time the agent reads some information from the XML file, a corresponding DIPPER update rule is triggered and the current information state is updated accordingly. Multiple levels of parsing are required and are performed using Prolog clauses. A second OAA agent appends the current information state values to the file that will finally contain the sequence of information states. The Communicator 2001 corpus contains evaluation information, e.g., about duration, task completion, etc., and therefore can be used for Reinforcement Learning whereas the 2000 collection lacks this additional information. However, the Communicator 2000 corpus comprises a large amount of data and it could be useful for other types of learning, e.g., rule-induction using RIPPER, or memory-based learning.

Experiments using the annotated Communicator data for learning dialogue strategies will be carried out at UEDIN in late 2004.

Chapter 3

WoZ-based Evaluation

As indicated in the introduction, spoken dialogue systems are complex and component level testing alone is not sufficient to fully categorise overall performance. In the early stages of development when complete working systems are not available, much can be learnt from examining simulated interactions between a human user and a human wizard pretending to take the role of the machine. These so-called Wizard of Oz (WoZ) collections can provide data for a variety of purposes:

- audio data and transcriptions for evaluating and refining acoustic and language models needed for speech recognition and understanding
- information on many aspects of dialogue design through the analysis of dialogue acts, response planning and generation, error recovery strategies, etc.
- annotated data for statistical dialogue modelling e.g., learning dialogue state transition probabilities, estimating user models and optimising dialogue strategies.

In this section, current and planned WoZ data collection activities within TALK are described. In section 3.1 conventional human-human studies are described. In these studies, the user and wizard can hear each other and recognition/understanding errors are rare. Hence, these studies are mostly useful for investigating information presentation issues. In order to investigate issues relating to the input side, including strategies for handling errors, then the effects of the ASR/speech understanding components must be simulated. Section 3.2 describes work done in that area within TALK. Finally, section 3.3 summarises the data collected so far.

3.1 Human-human Data Collection

There have been two human-human WoZ data collections. The first was conducted within the tourist information domain (part of the in-car scenario, and the second was conducted within the MP3 domain (relevant to both the in-car and the in-house scenarios).

3.1.1 Tourist Information Domain

The tourist information collection was part of the SACTI-1 (Simulated ASR-Channel – Tourist Information) described in section 3.2. This was a speech-only collection in which user's were given a map and asked to perform a number of tasks such as finding a hotel within a particular price range, or finding a restaurant of a particular type. An example map is shown in figure 3.1. This data was collected primarily as a contrast with the Simulated ASR-Channel data.

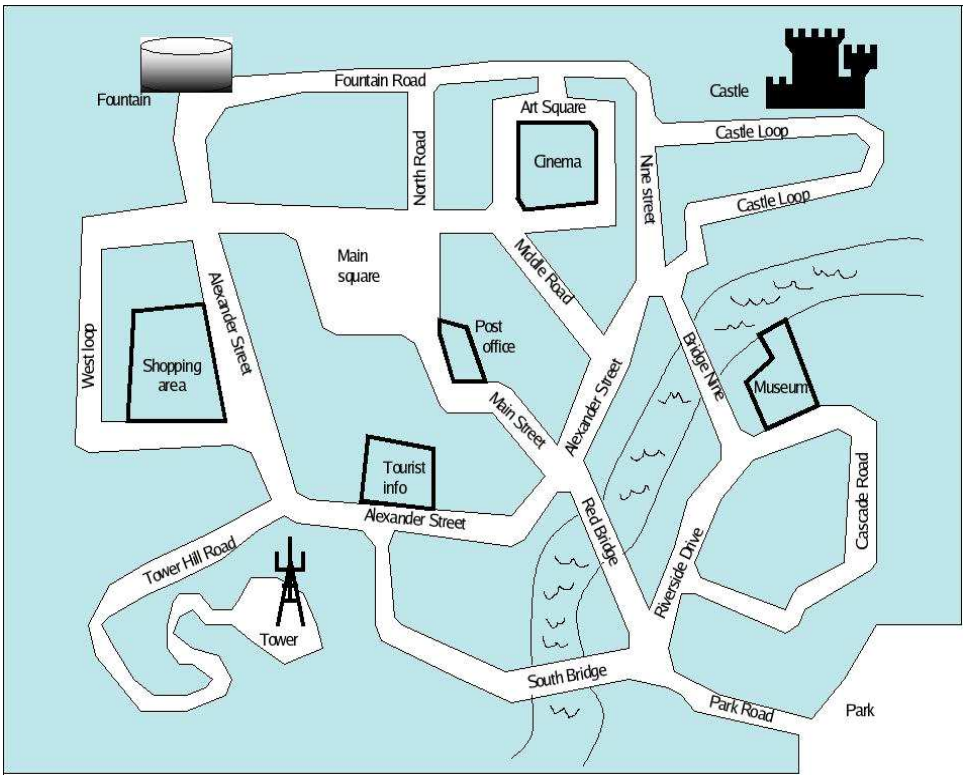

Figure 3.1: The map used for the tourist information data collection (SACTI)

3.1.2 MP3 Player Domain

A preliminary data collection was performed in German for the MP3 Player domain with the primary purpose of identifying different human speech presentation strategies for long lists as well as to elucidate phenomena to target in future experiments (see section 2.3). In this speech-only collection, wizards played the role of an MP3 player and were given access to a database of information (but not actual music) of more than 150,000 music albums (almost 1 million songs), extracted from the FreeDB database.[1] Figure 3.2 shows the experimental setup for our preliminary data collection and figure 3.3 shows an example screen shot for our music database as it is presented to the wizard. Subjects were given a set of predefined tasks and were told to accomplish them by using a speech-interface MP3 player. Tasks included such

[1] Freely available at http://www.freedb.org

things as playing songs/albums and building playlists, where the subject was given varying amounts of information to help them find/decide on which song to play or add to the playlist.

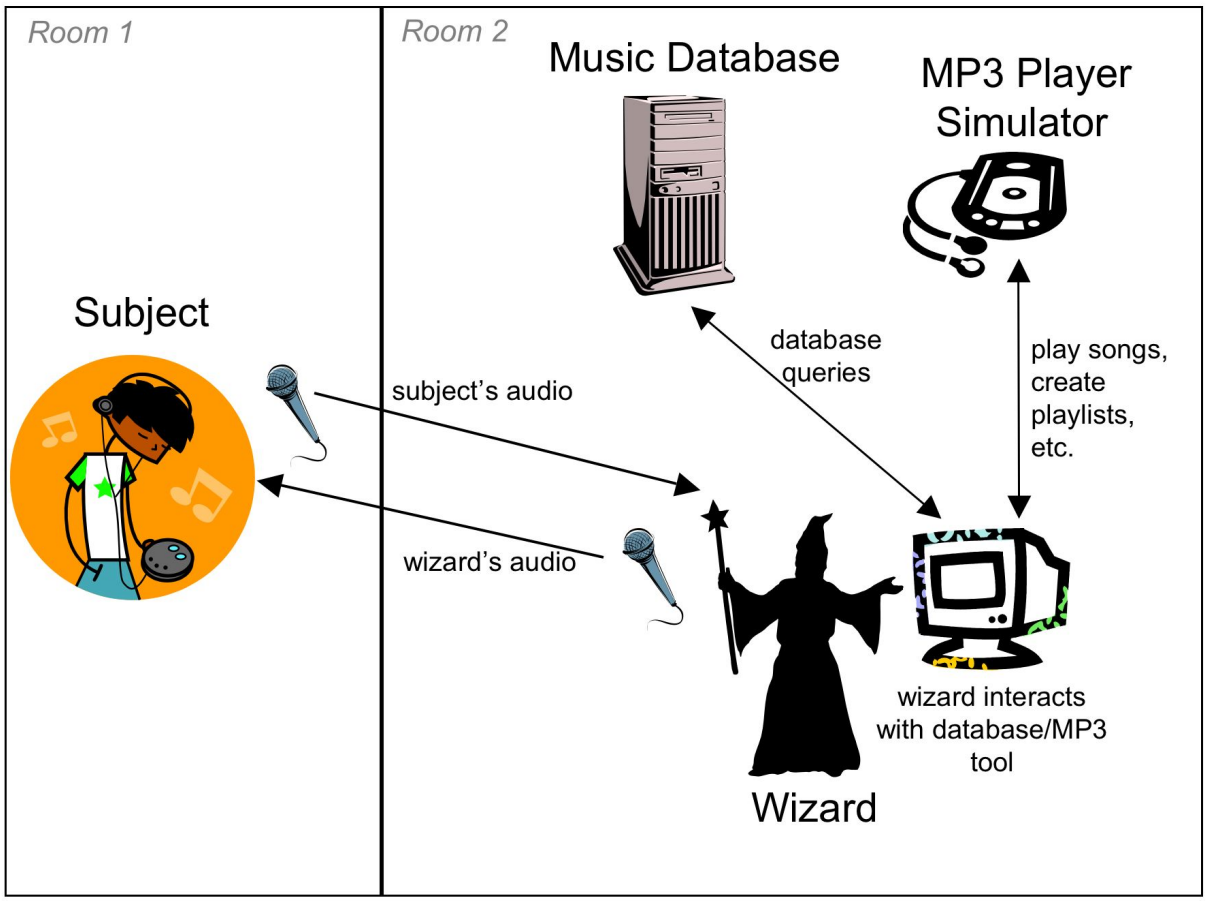

Figure 3.2: Experimental setup for the MP3 player domain.

The MP3 player domain is an interesting one from the point of view of presentation. Because we have access to such a large (and realistic) database of album and song information, subject queries can return thousands of matches, which then need to be communicated somehow by the system (wizard). A preliminary (audio) analysis of the data has revealed such human presentation techniques as presenting the entire list, asking if the list should be read, reporting just the number of results, and on-the-fly clustering of results and reporting on the cluster information.

The data from this collection is currently being transcribed[2] and will be annotated for (among other things) different presentation strategies for reporting query results.

[2]Hence the estimates in table 3.2, extrapolated from the sessions which have already been transcribed

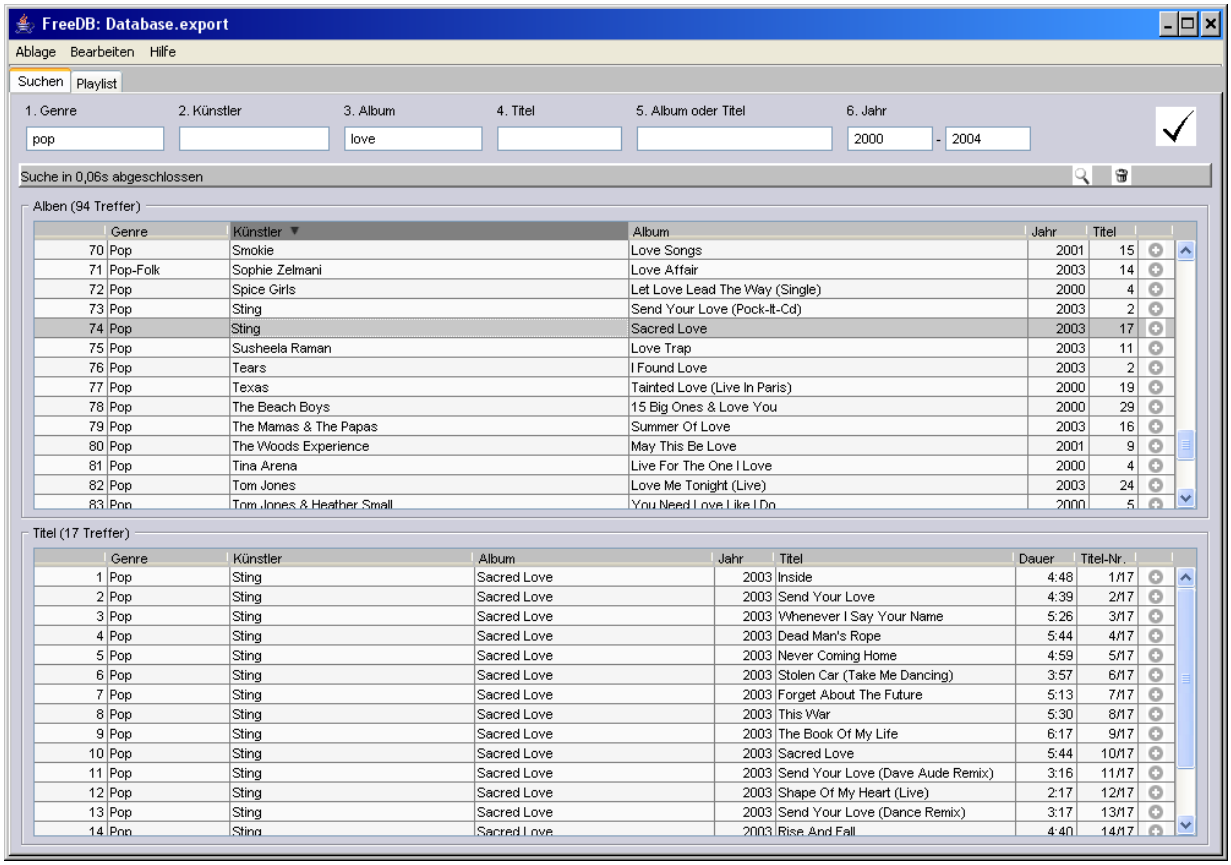

Figure 3.3: Screenshot from the FreeDB-based database application, as seen by the wizard.

3.2 Human-simulated-machine Data Collection

It is well known that end-pointing errors and speech recognition errors have a significant impact on the nature of human-machine dialogs making them very different to natural human-human dialogues. When collecting such WoZ data for research in dialogue management issues, it is therefore desirable to make the data as compatible with human-machine dialogues as possible. To achieve this a novel data collection framework has been developed which incorporates a simulated ASR channel.

The framework is based on a conventional “Wizard of Oz” set-up but has been modified as summarised in Figure 3.4. Two experimental participants, the “subject” and the “wizard” communicate via a simulated ASR channel. The participants are located in different rooms and cannot see each other.

The speech of both participants is end-pointed (i.e., segmented into utterances for performing recognition) using a standard energy-based end-pointer. The end-pointer is used to determine what wizard speech to play to the user, and what user speech to play to the typist. The end-pointing happens in just under real-time. The end-pointed utterances are saved for future analysis.

The subject can hear the wizard directly. However, the wizard cannot hear the subject; rather, both participants are told that the subject is speaking to a speech recogniser, which will take its best guess of what the subject says, and display it on a screen in front of the wizard.

When the system is busy and not listening to a participant, they hear a “tick-tock” sound. A turn-taking

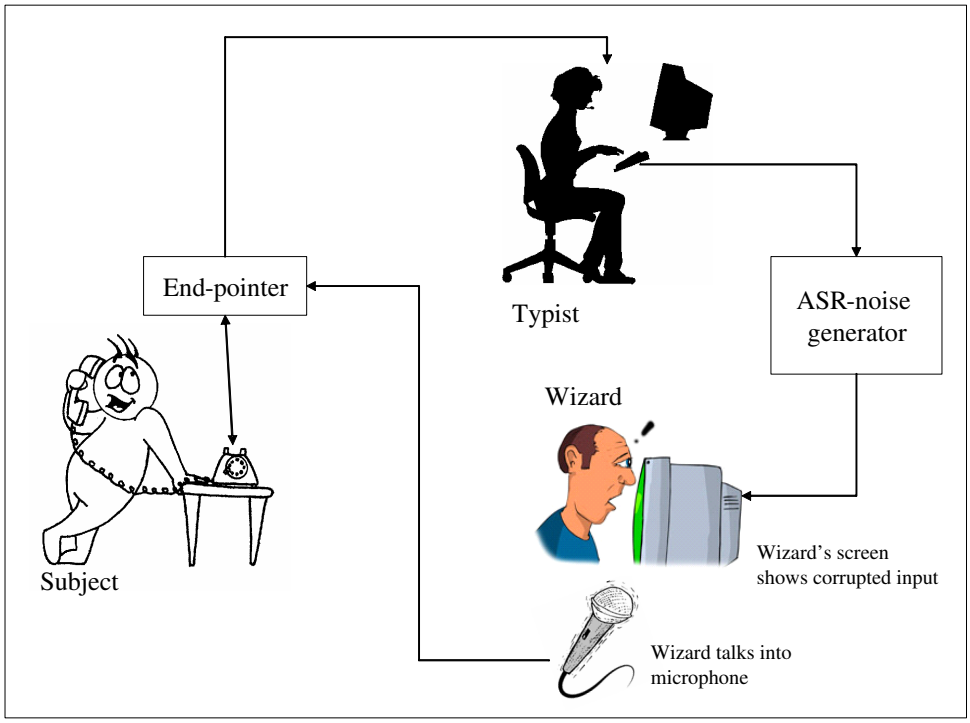

Figure 3.4: The WoZ Collection System with Endpointing and Simulated ASR Errors

model patterned after typical Human-Computer turn-taking models is used in which the user may "barge-in" over (interrupt) the wizard, but the wizard may not interrupt the user. In reality, the subject is speaking to a typist, who quickly transcribes the user's utterance. This transcription is passed to a system which simulates ASR errors, the output of which is displayed to the wizard. A state table describing these interactions is shown in figure 3.5.

The ASR component takes the utterance transcribed by the typist and introduces ASR-like errors in the text. It does this by first mapping the typed transcription into phones, then applying a phone confusion matrix, then "re-recognising" the speech using a task-dependent statistical language model. The aim of using this simulated ASR channel rather than a real system is twofold. First, the simulation allows the ASR error rate to be varied, allowing a range of system conditions to be evaluated. Second, using a simulated system at a target error rate is much quicker than building and running a real ASR system. In practice, error rates between 0% word error rate (WER) and 60% WER were used.

A detailed description of this collection framework is given in [SWY04], and a preliminary analysis of the SACTI-1 data is given in [WY04].

The SACTI data set is in two parts. SACTI-1 is speech only whereas for SACTI-2, the interface was extended to allow both the user and the wizard to click on locations on the map. To make this more useful, the wizard could also add the locations of all the hotels, restaurants, etc in the town. A screen shot of the enhanced map is shown in figure 3.6.

All of the SACTI data has been transcribed, and the turns marked. In addition, the SACTI-1 data has been annotated as described in [WY04] and the SACTI-2 data has the location of user and wizard clicks marked, and the wizard's use of display buttons. All annotations were done using ANVIL and are consistent with

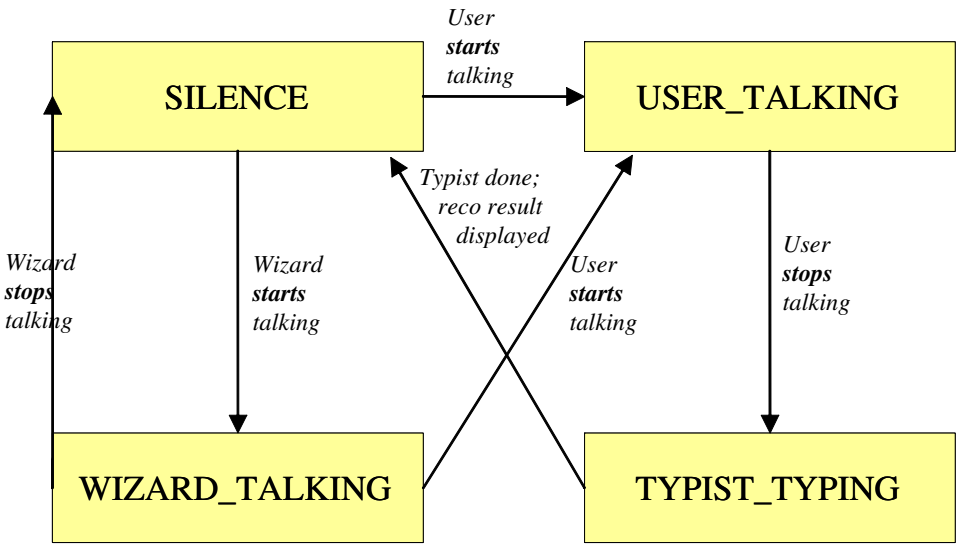

Figure 3.5: The State Transition Table for the Simulated ASR Collection Framework

the NXT annotation standard being adopted by TALK.

3.3 Summary of Data Collected to Date

This section provides various statistics describing the WoZ data collections performed within the first 9 months of the TALK project. Table 3.1 characterises the main characteristics of the data sets and Table 3.2 provides statistics on the quantity of data collected.

Name	Type	Domain	Input Modes	Output Modes	Language
SACTI-1	h-h	Tourist Info	Speech Only	Speech + Map display	English
SACTI-1	h-sm	Tourist Info	Speech Only	Speech + Map display	English
SACTI-2	h-sm	Tourist Info	Clicks/Buttons	Speech + Map display	English
MP3	h-h	MP3 Player	Speech Only	Speech Only	German

Table 3.1: Summary of Data Collections (h-h = human-human, h-sm = human-simulated-machine); all collections include speech input.

Figure 3.6: The SACTI Wizard's map enhanced to allow simple multimodal input/output

Name	#Dialogs	#Wizards	#Users	#Turns	#Words
SACTI-1	168	14	42	6113	75096
SACTI-2	179	6	36	4972	50727
MP3	24	2	24	∼2600	∼23000

Table 3.2: Data Collection Statistics.

Chapter 4

System Level Evaluation

This chapter will very briefly identify the working systems that we are using to progress the research. Each will be presented in terms of the research questions that they will address.

First, we identify design requirements on the experimental systems, which are intended to make the data collected useful across the project.

4.1 Design Requirements on the Experimental Systems

To maximize the use of data across experiments, we will determine a core set of data that will be collected in all experiments in common formats. This applies in particular to the system level experiments. Such data can include

- available environmental factors (e.g., performance in a primary task such as driving)
- raw data (audio signals (waveform) and gesture data (time-stamped click coordinates), graphical output)
- recognizer output (lattices or n-best for ASR and reference objects in gesture resolution)
- Interpretation and Generation results and processing details (including proto-content)
- full Information State and update rules fired
- application calls and returned data (e.g., database lookups)

In addition, we will determine which meta information such as data about subjects (e.g. age, gender), task completion, number of turns, number of times the user asked for help, and user satisfaction (from a questionnaire) we need to collect in a common format.

4.2 Research Systems

We now turn to the details of the individual dialogue systems that will be used for experiments in the TALK project.

4.2.1 The USAAR/DFKI System

USAAR and DFKI are building a dialogue system for the MP3 domain which will be used to encode and evaluate the types of multimodal presentation strategies observed from our WoZ data collections (see section 3.1.2). Of particular focus will be presentation strategies for long lists of results, which is relevant not only in the MP3 domain, but also in other domains that support database lookup (e.g., flight scheduling). The system will support basic MP3 player functions, including playlists as well as provide access to the FreeDB-based music information database described above. This system will also be the basis for the in-car demonstration system in month 18 (Task 5.2, Deliverable 5.2).

Extended versions of the system will add further domains suitable for the in-car scenario which will be determined in collaboration with Bosch and BMW. Relevant research questions will include presentation strategies for picking up interrupted tasks after varying amounts of time and how to vary presentations depending on user attention. User attention will be manipulated through a primary task. BMW has supplied software for a standard primary task in simulated car environments, the Lane Changing Task (LCT).

4.2.2 The UEDIN/UCAM Dipper System

A dialogue system built around the DIPPER dialogue manager [BKLO03] will be used to address issues in the evaluation of hand-coded versus learned dialogue strategies (see section 2.4). DIPPER allows us to rapidly build a baseline system for such experiments. This system will initially be used to conduct information-seeking dialogues with a user (e.g. find a particular hotel or restaurant), using hand coded dialogue strategies (e.g. always use implicit confirmation, except when ASR confidence is below 50%, then use explicit confirmation). We will then modify the DIPPER dialogue manager so that it can consult learned strategies (for example strategies learned from the 2000 and 2001 Communicator data), based on its current information state, and then execute dialogue actions from those strategies. This will allow us to compare hand-coded against learned strategies within the same system (i.e. the speech-synthesizer, recognizer, GUI, etc. will all remain fixed). The other core baseline dialogue system system components (communicating via OAA) are the Festival speech synthesizer and the ATK speech recognizer. A later version of the system will include a clickable map.

4.2.3 The Gothenburg In-Home Systems

UGOT are developing a research system in the in-home domain. Applications include MP3 player, Video player, X10 lights, and Agenda. In addition, an experimental multimodal map application is being developed. As a basis for implementation, some simple experiments with human-human dialogue will be conducted in the MP3 domain. As yet, there are no detailed plans for other experiments.

4.2.4 The Linguamatics In-home System

The Linguamatics System [MB03] provides a small-footprint, reconfigurable multi-modal system. There are two key ontology-based components: the interaction manager which decides on the next turn, and the interpretation component which analyses the input according to supplied ontologies. Ontologies have already been developed for home information and control, and the system has been installed at Loughbor-

ough University and linked to a home simulator. The Ergonomics and Safety Research Institute (ESRI) at Loughborough plans to test the usability of the home system with a variety of potential users.

As part of WP2, ontologies will be developed for the in-car scenario to test the reconfigurability of the system. The interpretation component will be separately tested to see if different strategies can improve performance.

4.2.5 The Seville In-Home System

The Delfos NCL dialogue management system [AQ02] will be used to test and learn multimodal dialogue strategies. This baseline system will be extended to include the new functionality intended for the In-Home showcase, addressing the specific research issues mentioned in WP1, WP2 and WP3. In addition to the new functionality, the system will integrate a touch-screen with a clickable display (home map, devices, lists, etc).

In order to conduct WOZ experiments, a research platform simulating the system functionality will be developed. This platform will include ASR, TTS, home set-up, OAA, video recording and display, remote screen supervision.

The wizard will simulate the system intelligence, which will be limited to that of the real system. The user will be presented with a graphical display in the touch-screen, and a microphone, and will be requested to perform certain tasks. They will be instructed how to use the devices, but they will not be given any particular instructions as to how to proceed. The users verbal input will be processed by the ASR, and the wizard will only see the ASR output, as well as any simultaneous screen input. The wizard will then interpret the output and provide the appropriate response, within a predefined set of possible actions.

The experiments will be video-recorded and all interaction between user and wizard will also be recorded.

Once preliminary results are obtained, the system will be modified according to those results and more testing will take place.

Chapter 5

Future Work

The proposed experiments and methods in TALK encompass three types of setups: component level, WoZ-based, and global system evaluations. They are described in sections 2 to 4. While these experiments concentrate on very specific research questions, some of the data collected can be re-used for other purposes and thus we plan to agree on a common core set of data and corresponding formats, see section 4.1. Finally, the following section 5.1 gives an overview over the schedule of the experiments to be conducted by each project partner.

5.1 Schedule of experiments

- USAAR/DFKI system
  - early 2005: first multimodal data collection, WoZ, for MP3 in-car
  - mid 2005: experiment with in-car baseline: evaluation of top-down vs. bottom-up vs. refinement strategies; parameter determination, optimal length of lists depending on various factors (Task 3.2)
  - late 2005: development of strategies for picking up interrupted tasks (Task 3.3)
  - late 2005: modality-specific realization strategies (Task 3.3)
  - early 2006: evaluation of strategies for task-switching situation (Task 3.3)
  - late 2006: evaluation of EIS-based (extended information state) systems vs. baseline (Task 3.3)
- UEDIN/UCAM system
  - late 2004: initial evaluation of learned dialogue strategies from the annotated Communicator corpus
  - late 2004: evaluation of statistical acoustic/language models in the Tourist Information Domain (Task 1.3)
  - early 2005: evaluation of confidence scoring metrics within MDP framework (Task 4.2)

  - mid 2005: comparing learned dialogue strategies with hand-coded strategies from the baseline system (Task 4.2)
  - mid-late 2005: context sensitive speech recognition versus general language models (Task 2.1d)
  - late 2005: evaluation of integrated multimodal statistical LMs by contrasting with separate monomodal input streams (Task 1.4)
  - late 2005: evaluation of user-models by comparing statistics with held-out SACTI data (Task 4.2)
  - mid 2006: evaluation of "voice programming" system (Task 2.3)
  - mid-late 2006: evaluation of POMDP-based systems compared with MDP-based, RL, and hand-coded dialogue strategies (Task 4.4)

- UGOT
  - will test components as described in the workplan for WP1.
- LING will perform component testing in WP2 tasks:
  - early 2005: usability testing (external to TALK) by University of Loughborough
  - early 2005: testing reconfigurability with ontologies for in-car domain (Task 2.1)
  - mid 2005: testing of interpretation routines (Task 2.1)
  - early 2006: testing of task plug-and-play (Task 2.2)
- USE
  - Late 2004: Set up for the WOZ experiments will be completed and tasks designed.
  - Late 2004-Early 2005: WOZ experiments will be carried out.
  - Mid 2005: Results will be analysed and integrated in the baseline system

5.2 Timeline

This overview summarizes the planned experiments in a timeline to give a different perspective. The items are abbreviated to allow a one-glance view, please refer to the previous section for full descriptions.

2004	
Late 2004	UEDIN/UCAM: initial evaluation of learned dialogue strategies, evaluation of statistical acoustic/language models

2005	
Early 2005	USAAR/DFKI: first multimodal data collection UEDIN/UCAM: evaluation of confidence scoring metrics LING: usability testing, testing reconfigurability with ontologies USE: WOZ experiments
Mid 2005	USAAR/DFKI: experiment with in-car baseline UEDIN/UCAM: comparing learned with hand-coded dialogue strategies LING: testing of interpretation routines
Late 2005	USAAR/DFKI: interrupted tasks, modality-specific realization strategies UEDIN/UCAM: context-sensitive speech recognition, multimodal statistical LMs, user-models

2006	
Early 2006	USAAR/DFKI: task-switching LING: task plug-and-play
Mid 2006	UEDIN/UCAM: voice programming
Late 2006	USAAR/DFKI: final EIS-based vs. base-line system UEDIN/UCAM: POMDP-based strategies

Table 5.1: Timeline of planned experiments.

Bibliography

[AQ02] J. Gabriel Amores and Jose F. Quesada. Cooperation and collaboration in natural command language dialogues. In *Proceedings of EDILOG*, 2002.

[BKLO03] Johan Bos, Ewan Klein, Oliver Lemon, and Tetsushi Oka. DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. In *4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124, Sapporo, 2003.

[CDE+99] Lynne Cahill, Christy Doran, Roger Evans, Chris Mellish, Daniel Paiva, Mike Reape, Donia Scott, and Neil Tipper. In Search of a Reference Architecture for NLG Systems. In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 77–85, Toulouse, 1999.

[CM01] Adam Cheyer and David Martin. The Open Agent Architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1/2):143–148, 2001.

[DBB+94] D Dahl, M Bates, M Brown, WM Fisher, K Hunicke-Smith, DS Pallet, C Pao, A Rudnicky, and E Shriberg. Expanding the scope of the atis task: the atis-3 corpus. In *Proc Human Language Technology Workshop*, pages 43–48, Plainsboro, NJ, 1994.

[KKERK03] Ivana Kruijff-Korbayova, Stina Ericsson, Kepa J. Rodriguez, and Elena Karagjosova. Producing contextually appropriate intonation in an information-state based dialogue system. In *Proceedings of EACL 2003*, pages 227–234, Budapest, Hungary, 2003.

[MB03] David Milward and Martin Beveridge. Ontology-based dialogue systems. In *Proceedings of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco, Mexico, 2003.

[MFLW04] Johanna Moore, Mary Ellen Foster, Oliver Lemon, and Michael White. Generating tailored, comparative descriptions in spoken dialogue. In *The 17th International FLAIRS Conference (Florida Artifical Intelligence Research Society)*, 2004.

[Ovi99] Sharon L. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 1999.

[SWY04] MN Stuttle, JD Williams, and SJ Young. A framework for dialog systems data collection using a simulated asr channel. In *ICSLP 2004*, Jeju, Korea, 2004.

[SY01] Konrad Scheffler and Steve Young. Corpus-based dialogue simulation for automatic strategy learning and evaluation. In *Proceedings of NAACL Workshop on Adaptation in Dialogue Systems*, 2001.

[WAB+01] M Walker, J Aberdeen, J Boland, E Bratt, J Garofolo, L Hirschman, A Le, S Lee, S Narayanan, K Papineni, B Pellom, B Polifroni, A Potamianos, P Prabhu, A Rudnicky, G Sanders, S Seneff, D Stallard, and S Whittaker. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Eurospeech 2001*, Aalborg, Scandinavia, 2001.

[WY04] JD Williams and SJ Young. Characterising task-oriented dialog using a simulated asr channel. In *ICSLP 2004*, Jeju, Korea, 2004.

[You00] SJ Young. Probabilistic methods in spoken dialogue systems. *Philosophical Transactions of the Royal Society (Series A)*, 358(1769):1389–1402, 2000.