Journal of Computing and Information Technology - CIT 15, 2007, 2, 111–121
doi:10.2498/cit.1000795

A Dialogue-management Evaluation Study

Melita Hajdinjak and France Mihelič

Faculty of Electrical Engineering, University of Ljubljana, Slovenia

We present a highly portable and cooperative dialogue-manager component of a developing, Slovenian and Croatian spoken dialogue system for weather-information retrieval. In order to evaluate the performance of this component, two Wizard-of-Oz experiments were performed. The only difference between the two experiment settings was in the dialogue-management manner, i.e., while in the first experiment dialogue management was performed by a human, the wizard, in the second experiment it was performed by the newly-implemented dialogue-manager component. The data from both Wizard-of-Oz experiments was evaluated with the PARADISE evaluation framework,which was proposed as a potential general methodology for evaluating and comparing different versions of spoken-language dialogue systems. The study ascertains a remarkable difference in the performance functions when taking different satisfaction-measure sums, or even individual scores as the target to be predicted, it introduces the dialogue costs *database parameters*, and it confirms the dialogue manager's cooperativity subject to the incorporated knowledge representation.

Keywords: dialogue system, dialogue management, conversational game theory, Wizard-of-Oz experiment, dialogue-system evaluation, PARADISE framework

1. Introduction

There has recently been a great deal of interest in developing dialogue systems for accessing information sources through the telephone network [16, 24] or the internet [4] using spoken or written natural language. However, the central module of any natural-language dialogue system (Figure 1) is the dialogue manager, which plays the role of an intermediate agent between the user and the knowledge source. The dialogue manager operates on a meaning representation, modeling what the user has written or what the speech-recognition module has recognized, and its overall goal is to take an active role in directing the dialogue flow toward a successful conclusion for the user.

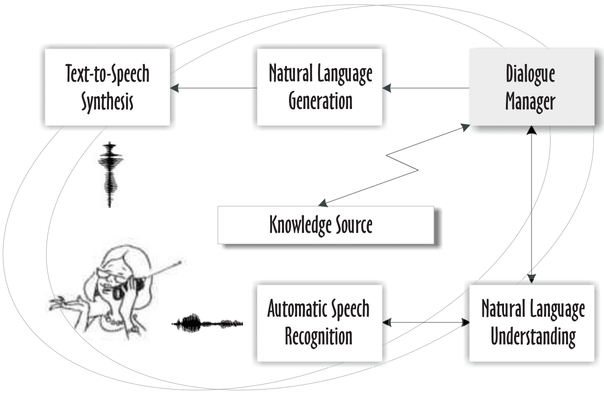
Text-to-Speech Synthesis
Natural Language Generation
Dialogue Manager
Knowledge Source
Automatic Speech Recognition
Natural Language Understanding

Figure 1. Natural-language dialogue system.

Most of the natural-language dialogue systems constructed to date use the *slot-filling* (*frame-based*) approach to dialogue management [6], where the dialogue manager responds to user's queries with a sequence of clarifications to obtain enough information in order to perform a specific action. In this approach the task and the dialogue strategy are separated; the task is to fill the slots, which can be achieved using various dialogue strategies. These strategies are independent of the contents of the slots, this is why they can be reused when porting the system to a new domain. In Section 2, a highly portable, slot-filling dialogue-manager component [7] of a developing, bilingual-spoken dialogue system for weather-information retrieval [25] is presented. The underlying dialogue strategy is modeled using *conversational game theory*, which represents a line of research [18, 11, 13, 17] where dialogues consist of exchanges, called *conversational games*.

In order to evaluate the performance of spoken-language dialogue systems, Walker et al. [21] proposed PARADISE (PARAdigm for DIalogue System Evaluation), a framework that models user satisfaction as a linear combination of measures reflecting *task success* and *dialogue costs*. Some important PARADISE details and issues were, however, highlighted by Hajdinjak and Mihelič [9]. Applying PARADISE to dialogue data requires the dialogue corpora to be collected via controlled experiments during which users subjectively rate their satisfaction. Therefore, in order to evaluate the dialogue manager of the developing, bilingual-spoken dialogue system [25], two Wizard-of-Oz (WOZ) experiments [8] were conducted. While in the first WOZ experiment dialogue management was still one of the tasks of the wizard, in the second WOZ experiment it was performed by the dialogue-manager component. Both experiment settings are described in Section 3.

We claim that the influence of automatic speech recognition hinders the other parameters from showing significance when evaluating the performance of the dialogue-management process. Therefore, in both WOZ systems, which were carried out in order to evaluate dialogue management, automatic speech understanding was not performed. In Section 4, the application of PARADISE to the data from both WOZ experiments is detailed, and interesting evaluation results are given.

2. Dialogue Management

In the developing, Slovenian and Croatian spoken dialogue system for weather-information retrieval [25, 7] the slot-filling approach [6] to dialogue management was used, and because of the relative simplicity (i.e., a property common to almost all information-providing domains) of the weather domain only three different types of slots, i.e., *location*, *time*, and *information*, were defined. Moreover, a special knowledge representation [7], which is consistently flexible in directing the user to select relevant, available data, was incorporated into the dialogue-management process.

The dialogue strategy was modeled using *conversational game theory* [18, 11, 13, 17], where conversations are structured on two functional levels, i.e., *conversational games* and *conversational moves*. The level of conversational games is associated with mutually understood conversational goals, such as obtaining information or getting the conversational partner to perform a specific action. They are made up of sets of utterances starting with an initiation and encompassing all utterances up until the purpose of the conversational game has been either fulfilled or abandoned, and are analysed as conversational moves where a move is an utterance, a partial utterance, or a group of utterances that convey the same specific intent, such as instructing or requesting a clarification. However, a theoretical account of dialogue, where conversational moves are viewed as objects that update, revise, and align the informational states of the conversational partners, was promoted in the TRINDI project [14].

Note, the construction of the dialogue manager was guided by the evaluation outcomes (Section 4) of the data from the first WOZ experiment (Subsection 3.1).

2.1. Modeling the Dialogue Strategy

Obviously, the ability to employ a rich set of conversational strategies greatly influences the usability and ultimately the success of natural-language dialogue systems. Therefore, conversational games that encompass not only grounding behaviours, e.g., confirmations and disambiguation, and turn-taking behaviour, but also the ability to handle requests for help and providing context-specific help messages, for repeating the last statement, suspending the dialogue, and re-establishing the context were defined. These definitions were made in agreement with the findings of the first WOZ experiment [8] and according to the coding system [1], applied to a corpus of spontaneous task-oriented spoken dialogues.

Consequently, we decided to distinguish three basic types of conversational moves:

- *initiating moves* occur at the beginning of a game, where they introduce a new discourse purpose into the dialogue;
- *response moves* occur within games, after an initiation and serve to fulfill the expectations set up within the game;
- *ready moves* occur after a game's closing and prepare the conversation for a new game to be initiated.

Furthermore, we came to the conclusion that an extension of the set of conversational games implemented in the TRINDI project [14] is needed to enable greater portability and/or greater cooperativity of the dialogue system. We defined 11 fundamental initiating moves;

- GREET Indicates a greeting.
- INDECIPHERABLE Indicates a user's query that was indecipherable to the system.
- PARDON Indicates an asking for repetition of the last query.
- HELP Indicates a user's appeal for help.
- TIMEOUT Indicates a system's belief that the user didn't say anything in the allotted time.
- INTERRUPT Indicates a user's interruption of playing an information-providing game.
- ALIGN Indicates a user's checking to see if the system's understanding is in accordance with his/her understanding.
- CHECK Indicates a system's question about something that it believes it already knows the answer to, but is not absolutely certain. These moves cover past dialogue events.
- END Indicates a user's decision to end the conversation.
- QUERY-YN Indicates any question that takes yes or no as the answer and does not count as a CHECK or an ALIGN move.
- QUERY-WR Mostly indicates a wh-question, a request for certain information or additional data.

and 3 slot-related initiating moves;

- QUERY-WI Refers to the slot *information* and indicates a user's request to list the types of information that the system is able to provide.
- QUERY-WL Refers to the slot *location* and indicates a user's request to list the spatial data for which the system is able to provide the requested information.
- QUERY-WT Refers to the slot *time* and indicates a user's request to list the temporal data for which the system is able to provide the requested information.

7 fundamental response moves;

- ACKNOWLEDGE Indicates a verbal response that minimally shows that the move was understood and/or accepted.
- REPLY-HELP Indicates a system's reply to a HELP move.
- REPLY-TIMEOUT Indicates a system's reply to a TIMEOUT move.
- REPLY-Y Indicates a reply with yes to any query with a yes-no possible answer (i.e., QUERY-YN, CHECK, ALIGN).
- REPLY-N Indicates a reply with no to any query with a yes-no possible answer (i.e., QUERY-YN, CHECK).
- REPLY-MOD Indicates a reply with a correction to a yes-no possible answer (i.e., QUERY-YN, CHECK, ALIGN).
- REPLY-WR Indicates a reply to a QUERY-WR move.

and 3 slot-related response moves;

- REPLY-WI Indicates a system's response to a QUERY-WI move.
- REPLY-WL Indicates a system's response to a QUERY-WL move.
- REPLY-WT Indicates a system's response to a QUERY-WT move.

one ready move;

- READY Indicates that the previous game has just been completed and a new game is about to begin.

With respect to the extended set of conversational moves, 15 conversational games (Figure 2), which correspond to the initiating moves and the ready move, respectively, were implemented. These games are formalized as *recursive transition networks*, i.e., diagrams consisting of paths that may be followed and of operations along these paths that must be carried out, that permit arbitrary nesting, i.e., they enable any conversational game to occur at any point within any other conversational game as soon as one game is initiated to serve the larger goal of a game that has been initiated before.

Note, slot-related conversational games (i.e., QUERY-WI GAME, QUERY-WL GAME, and QUERY-WT GAME), each referring to one of the defined slots, have not been used before. Such slot-related games, on the one hand, offer the advantage of representing the human-computer dialogue flow in a more structured way, and, on the other hand, enable users to ask for available data. This is very important in information-providing dialogue systems where the need to inform the user about the scope of the system's knowledge is one of the most critical aspects [24], in particular when relying on a sparse and dynamical information source with a time-dependent data structure.

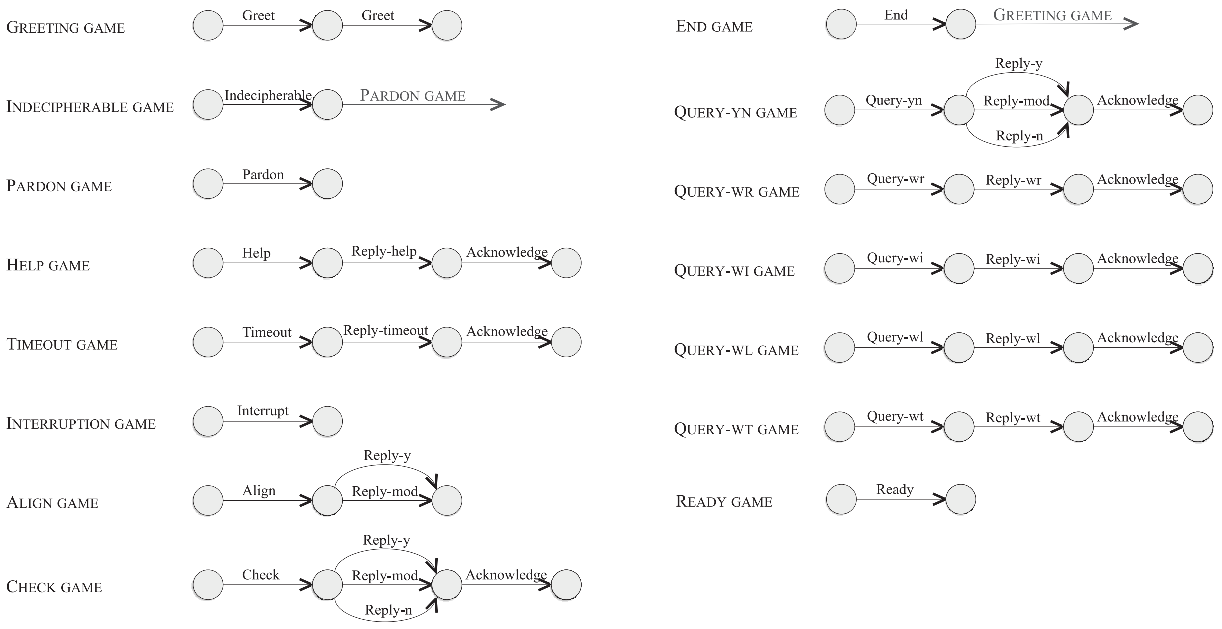

Figure 2. Implemented conversational games.

The dialogue in Table 1, generated using the system's dialogue-manager component and natural-language generation, shows a number of employed conversational strategies; GREETING GAME, initiated by S1 and S10; QUERY-WR GAME, initiated by U1, U2, U5, and U7; CHECK

S1:	*Hello. The dialogue system for weather-forecast information speaking. Can I help you?*
U1:	Yes, I would like the weather for Slovenia tomorrow.
S2:	*In Slovenia tomorrow - partly cloudy, the rainfalls and the snowfalls will stop. Is there something else?*
U2:	Yes. What about the weather in the mountains?
S3:	*Are you asking for tomorrow?*
U3:	...
S4:	*You haven't told me all the necessary information or I haven't understood it. If you have any other question, please ask, otherwise answer the following question: Are you asking for tomorrow?*
U4:	No, I would like the current weather please.
S5:	*Sorry, but I have no data about the current weather in the mountains. However, I have data about the current weather on Kredarica, Vogel, Kanin, Krvavec, and Predel. I can also give you the weather forecast in the mountains for today and for tomorrow.*
U5:	Tell me for Vogel please.
S6:	*Are you asking about the current weather?*
U6:	Yes.
S7:	*On Vogel at this moment - fog. Anything else?*
U7:	Yes, tell me the air temperature and the depth of the snow on Vogel please.
S8:	*On Vogel at this moment - the air temperature is -7 degrees Celsius, the depth of the snow is 180 cm. Do you have any other question?*
U8:	Could you tell me what kind of information you have for Ljubljana at this moment?
S9:	*For Ljubljana at this moment I have information about the weather, the humidity, the air temperature, the wind speed, and the wind direction. Is there anything I can inform you about?*
U9:	No. That's all.
S10:	*Thank you. Goodbye.*

Table 1. Slovene-English translation of an example dialogue.

GAME, initiated by S3 and S6; TIMEOUT GAME, initiated by U3 when the user did not respond in the allotted time; QUERY-WI GAME, initiated by U8; END GAME, initiated by U9. Indispensable in re-establishing the context, however, is the conversational game CHECK GAME, which covers past dialogue events. In this game, instead of making definite conclusions, the system requests the user to confirm the information that it already assumes to be true. The REPLY-WR move S5, on the other hand, shows how the system directs the user to select relevant, available data when not being able to give the answer to his/her explicit request.

3. Wizard-of-Oz Experiments

The construction of the dialogue manager (Section 2) began by conducting the WOZ experiment [5, 2]. In WOZ studies subjects are told to interact with a computer system, though in fact they are not. The system is at least partly simulated by a human, the wizard, with the consequence that the subjects can be given more freedom of expression or be constrained in more systematic ways than this is the case in already existing dialogue systems. Since the dialogues collected during such an experiment reflect the language that would be attempted when communicating with a computer system, the WOZ experiment proves successful in the early stages of the construction of natural-language dialogue systems.

Hence, while the aim of the first WOZ experiment was, first of all, to collect human-computer data, the aim of the second WOZ experiment was to evaluate the newly-implemented dialogue-manager component.

3.1. First WOZ Experiment

The task of the wizard in the first WOZ experiment [8] was to simulate Slovenian speech understanding (i.e., speech recognition and natural-language understanding) and dialogue management within the weather-information-providing, Slovenian and Croatian spoken dialogue system [25]. Since only Slovene users were involved into the experiment, Croatian speech understanding was not performed.

However, a total of 76 Slovene users (38 female, 38 male) were chosen to take part in the first WOZ experiment. They were given verbal instructions about the general functionality of the system and a sheet of paper containing a description of the tasks they were supposed to complete. The users had two scenarios to enact. The first task was to obtain a particular piece of weather-forecast information, such as the temperature in London or the weather forecast for Slovenia tomorrow, and the second task was a given situation, such as "You are planning a trip to... What are you interested in?", the aim of which was to stimulate the user to ask context-specific questions. After these two scenarios, and in contrast to previous experiments, users were given the freedom to ask additional questions.

In order to evaluate user satisfaction, users were given the user-satisfaction survey (Table 2) used within the PARADISE framework, which asks to specify the degree to which one agrees with several questions about the behaviour or the performance of the system (**TTS Performance**, **ASR Performance**, **Task Ease**, **Interaction**

1. *Was the system easy to understand?* (**TTS Performance**)
2. *Did the system understand what you said?* (**ASR Performance**)
3. *In this conversation, was it easy to find the message you wanted?* (**Task Ease**)
4. *Was the pace of the interaction with the system appropriate?* (**Interaction Pace**)
5. *In this conversation, did you know what you could say at each point of the dialogue?* (**User Expertise**)
6. *How often was the system sluggish and slow to reply to you?* (**System Response**)
7. *Did the system work the way you expected it to?* (**Expected Behaviour**)
8. *From your current experience with using the weather-information providing dialogue system, do you think you'd use the system regularly when you need information about the weather?* (**Future Use**)

Table 2. User-satisfaction survey used within the PARADISE framework.

Pace, **User Expertise**, **System Response**, **Expected Behaviour**, **Future Use**). The answers to the questions were based on a five-class ranking scale from 1, indicating strong disagreement, to 5, indicating strong agreement. A comprehensive **User Satisfaction** was then computed by summing each question's score, and thus ranged in value from a low of 8 to a high of 40. In the first WOZ experiment, the mean **User Satisfaction** value was 34.08, with a standard deviation of 5.07.

3.2. Second WOZ Experiment

In contrast to the first WOZ experiment, the task of the wizard in the second WOZ experiment was only to simulate Slovenian speech understanding. The wizard was sitting behind the interface of the newly-implemented dialogue manager and entered the meaning representation of what the user said. All the other components of the system remained the same.

A total of 68 Slovene users (29 female, 39 male) were chosen to take part in the second WOZ experiment. They were given the same instructions and the same user-satisfaction survey as in the first experiment. The mean **User Satisfaction** value was 31.96, with a standard deviation of 4.99.

Note, it was expected that in the second experiment the **User Satisfaction** value would be a bit worse (statistically significant with $p < 0.015$) since the wizard with her human-level intelligence should have been able to manage the dialogue better than the implemented dialogue-manager component.

4. Evaluation

In order to find the most significant parameters (i.e., *predictors*) of the dialogue manager's performance, the PARADISE framework [21] was used. It maintains that the system's primary objective is to maximize user satisfaction, and it derives a combined performance metric for a dialogue system as a weighted linear combination of *task-success measures* and *dialogue costs*. Consequently, applying PARADISE to dialogue data requires the dialogue corpora to be collected via controlled experiments during which users subjectively rate their satisfaction. In addition, the other parameters of the model of performance, i.e. the task-success measures and the dialogue costs, must be either automatically logged by the system or hand-labeled.

The PARADISE model of performance [22] posits that a performance function can then be derived by applying multivariate linear regression (MLR) with **User Satisfaction** as the dependent variable and task-success measures and dialogue costs as the independent variables:

$$Performance = (\alpha * \mathcal{N}(\kappa)) - \sum_{i=1}^{n} w_i * \mathcal{N}(c_i)$$

Here, α is the weight on the Kappa coefficient κ [3], which can be calculated from a confusion matrix that summarizes how well the dialogue system achieves the information requirements of particular tasks within the dialogue and measures task success, w_i are weights on the dialogue costs c_i, and $\mathcal{N}$ is a Z-score normalization function:

$$\mathcal{N}(x) = \frac{x - \overline{x_0}}{\sigma_{x_0}}$$

where $\overline{x_0}$ and σ_{x_0} are the mean value and the standard deviation for x, respectively, computed from the sample set of observations. The normalization function $\mathcal{N}$ guarantees that the weights directly indicate the relative contributions to the performance function, which can be used to predict **User Satisfaction**.

Because of the often reported finding [22, 12, 23, 15] that the mean concept accuracy, often referred to as the mean recognition score, is the exceptional predictor of a dialogue system's performance, we claim that the influence of speech recognition hinders the other parameters from showing significance when evaluating the performance of a specific component of a dialogue system. Therefore, in our WOZ experiments (Section 3), which were carried out in order to compare and to evaluate both dialogue-management manners, speech understanding was performed by a human wizard.

4.1. Selection of Regression Parameters

The selection of the regression parameters is of great importance and, therefore, has to be made on a thorough consideration. In order to compare the performance of both WOZ systems, 25 regression parameters were selected, i.e. the task-success measure

- **Kappa coefficient** (κ), reflecting both the wizard's typing errors and the unauthorized, mostly relevant changes in the meaning representations of user's utterances,

and the dialogue costs

- **Mean Elapsed Time** (MET), i.e. the mean elapsed time of user-initiated, information-providing conversational games (QUERY-WR GAMES and QUERY-YN GAMES) that occurred within the interaction,
- **Mean User Moves** (MUM), i.e. the mean number of conversational moves that the user needed to either fulfil or abandon the initiated information-providing games,
- **Task Completion** (Comp), i.e. the user's perception of completing the first task,
- **Number of User Initiatives** (NUI), i.e. the number of user's moves initiating information-providing games,
- **Mean Words per Turn** (MWT), i.e. the mean number of words per user's turn,
- **Mean Response Time** (MRT), i.e. the mean system-response time,
- **Number of Missing Responses** (NMR), i.e. the difference between the number of system's turns and the number of user's turns, which, on the one hand, reflects the number of user's TIMEOUT moves, and, on the other hand, his/her unreadiness to greet the system,
- **Number of Unsuitable Requests** (NUR) and **Unsuitable-Request Ratio** (URR), i.e. the number and the ratio of user's initiating moves that were out of context,
- **Number of Inappropriate Responses** (NIR) and **Inappropriate-Response Ratio** (IRR), i.e. the number and the ratio of contextually inappropriate system's responses, including PARDON moves,
- **Number of Errors** (Error), i.e. the number of system errors, including interruptions of the telephone connection, unsuitable natural-language sentences, and contradictory statements,
- **Number of Help Messages** (NHM) and **Help-Message Ratio** (HMR), i.e. the number and the ratio of system's REPLY-HELP and REPLY-TIMEOUT moves,
- **Number of Check Moves** (NCM) and **Check-Move Ratio** (CMR), i.e. the number and the ratio of system's CHECK moves,
- **Number of Given Data** (NGD) and **Given-Data Ratio** (GDR), i.e. the number and the ratio of system's information-providing moves,
- **Number of Relevant Data** (NRD) and **Relevant-Data Ratio** (RDR), i.e. the number and the ratio of system's moves directing the user to select relevant, available data,
- **Number of No Data** (NND) and **No-Data Ratio** (NDR), i.e. the number and the ratio of system's moves stating that the requested information is not available, and
- **Number of Abandoned Requests** (NAR) and **Abandoned-Request Ratio** (ARR), i.e. the number and the ratio of the information-providing games that were abandoned by the user.

Note, we considered both quantitative and proportional parameters in order to ascertain users' sensitivities.

All the mean values of the listed parameters are given in Table 3. Those that showed a significant change in value across both WOZ experiments are shaded grey and the corresponding p value [19] is given. The p value is a measure of how much evidence we have against the null hypotheses, i.e. the probability that our sample could have been drawn from the population(s) being tested (or that a more improbable sample could be drawn) given the assumption that the null hypothesis is true.

First, Table 3 says that MET and MUM were significantly greater (i.e. $p < 0.0005$ and $p < 0.05$, respectively) in the second WOZ experiment. Obviously, this was partly due to the conversational game CHECK GAME, which was implemented in the dialogue-manager component. Moreover, since the majority of the replies to CHECK moves contained less than three words, this dialogue strategy gave rise to a significantly lower ($p < 0.0005$) value of MWT. Nevertheless, the increase of MET was also influenced by the significantly longer ($p < 0.0005$) system's response times (MRT) in the second experiment, which was a reflection of the wizard's time-consuming typing of the meaning representations of users' utterances.

Second, an interesting finding is the relatively high negative correlation (i.e. -0.53 in the first experiment and -0.51 in the second experiment) between NUI and MUM, which reflects the users' ability to adapt to the system's behaviour and to learn how to more efficiently complete the tasks.

Third, special attention was given to the parameters NGD, GDR, NRD, RDR, NND, and NDR, which have not so far been reported in the literature as costs for user satisfaction. We

		WOZ1	WOZ2	p
task success	**Kappa koeficient** (κ)	0.94	0.98	
efficiency	**Mean Elapsed Time** (MET)*	13.76 s	17.39 s	0.000
costs	**Mean User Moves** (MUM)	1.48 s	1.68 s	0.047
	Task Completion (Comp)	0.97	0.96	
	Number of User Initiatives (NUI)	6.49	7.51	0.005
	Mean Words per Turn (MWT)	9.32 s	7.56 s	0.000
	Mean Response Time (MRT)	5.13 s	6.38 s	0.000
	Number of Missing Responses (NMR)	0.60	0.75	
	Number of Unsuitable Requests (NUR)	0.48	0.13	0.011
	Unsuitable-Request Ratio (URR)	0.08	0.02	
	Number of Inappropriate Responses (NIR)	0.41	0.90	0.009
	Inappropriate-Response Ratio (IRR)	0.04	0.06	
quality	**Number of Errors** (Error)	0.12	0.06	
	Number of Help Messages (NHM)	0.32	0.40	
costs	**Help-Message Ratio** (HMR)	0.03	0.03	
	Number of Check Moves (NCM)⋆	0	2.19	0.000
	Check-Move Ratio (CMR)⋆	0	0.16	0.000
	Number of Given Data (NGD)	4.07	4.35	
	Given-Data Ratio (GDR)	0.67	0.58	
	Number of Relevant Data (NRD)	0.70	2.06	0.000
	Relevant-Data Ratio (RDR)	0.10	0.28	0.005
	Number of No Data (NND)	1.67	0.94	0.000
	No-Data Ratio (NDR)	0.22	0.12	
	Number of Abandoned Requests (NAR)	0.05	0.16	
	Abandoned-Request Ratio (ARR)	0.01	0.02	
	User Satisfaction (US)	34.08	31.96	0.015

* Duration of the system's replies is not included.

⋆ In the first WOZ experiment, the wizard did not perform CHECK moves.

Table 3. Mean values of the selected parameters in the first (WOZ1) and the second (WOZ2) WOZ experiment.

will refer to them as *database parameters*. It has, however, been argued [22] that the database size might be a relevant predictor of performance. Indeed, in our experiments, relying on the extremely sparse and dynamical weather-information source [7] with a time-dependent data structure, it turned out that these parameters can play an important part in predicting the performance of a dialogue system, the performance of its specific components, and even in predicting individual user-satisfaction metrics (Subsection 4.3).

Another interesting finding is that, although in the first experiment the users were more comprehensive for quantitative database parameters (i.e. NGD, NRD, NND) than for proportional database parameters (i.e. GDR, RDR, NDR), in the second experiment it was just the other way round.

In addition, Table 3 says that in the second WOZ experiment NRD and RDR were almost three times greater than in the first experiment, which confirms the dialogue manager's ability to direct the user to select relevant, available data when his/her explicit request yields no information. Consequently, in the second WOZ experiment, NND was significantly lower ($p < 0.0005$).

4.2. Why Model the Sum of User-Satisfaction Scores

What if we want to evaluate a specific component of a dialogue system (e.g., automatic speech recognition or dialogue-manager performance)? In compliance with Hone and Graham's [10] observations, we argue that the approach of summing all the user-satisfaction scores can only be justified on the basis of evidence that all of he items are measuring the performance of the chosen dialogue-system component, otherwise the overall score is meaningless.

Indeed, our experiments showed a remarkable difference in the significance of the parameters when taking different satisfaction-measure sums or even individual scores as the target to be predicted (Subsection 4.3). Another interesting finding is that some individual user-satisfaction metrics, could not be well modeled.

4.3. Performance Function Results

As the target to be predicted we first took **User Satisfaction** (US) and afterwards the sum of those user-satisfaction values that (in our opinion) measured the dialogue manager's performance (DM), i.e. the sum of the user-satisfaction-survey scores assigned to the questions associated with **ASR Performance**, **Task Ease**, **System Response**, and **Expected Behaviour** (Table 2). The selected user-satisfaction values could all be well modeled and they were all under the influence of the dialogue-management manner.

Eliminating outliers, i.e. observations that lie at an abnormal distance from other values, is a common practice in multivariate linear regression [20]. In compliance with this practice, about 10% of the outliers in the data from both WOZ experiments were removed.

However, considering the data from the first WOZ experiment, backward elimination for $F_{out} = 2$ [19] gave the following performance equations:

$$\widehat{\mathcal{N}(US)} = -0.69\mathcal{N}(NND) - 0.16\mathcal{N}(NRD)$$

$$\begin{aligned}\widehat{\mathcal{N}(US)} &= -0.61\mathcal{N}(NND) + 0.21\mathcal{N}(Comp) \\ &= -0.16\mathcal{N}(NRD)\end{aligned}$$

with 58% (i.e., $R^2 = 0.58$) and 59% of the variance explained, respectively. To be able to observe the close similarity between these two equations, note that Comp was significant for US ($p < 0.02$), but removed by backward elimination.

In contrast, considering the data from the second WOZ experiment, backward elimination for $F_{out} = 2$ gave the following performance equations:

$$\begin{aligned}\widehat{\mathcal{N}(US)} &= -\,0.30\mathcal{N}(CMR) - 0.23\mathcal{N}(MET) \\ &\quad + 0.18\mathcal{N}(\kappa)\end{aligned}$$

$$\begin{aligned}\widehat{\mathcal{N}(DM)} &= -\,0.35\mathcal{N}(CMR) + 0.35\mathcal{N}(GDR) \\ &\quad + 0.35\mathcal{N}(\kappa) - 0.17\mathcal{N}(ARR)\end{aligned}$$

with 26% and 46% of the variance explained, respectively. Again, it is necessary to know that *MET* was significant for DM ($p < 0.02$) and that *GDR* and *ARR* were significant for US ($p < 0.04$), but all removed by backward elimination.

Let us compare both performance equations predicting DM, ascertain the effect of the dialogue-manager component. The first observation that we make is that none of the predictors is common to both performance equations. All the predictors from the first performance equation (i.e., NND, Comp, NRD) were insignificant ($p > 0.1$) for DM in the second experiment. On the other hand, the only predictor from the second performance equation that was significant for DM ($p < 0.004$) in the first experiment, but removed by backward elimination, was GDR.

Unlike the first performance equation with the database parameters NND and NRD as crucial negative predictors, the second performance equation clearly shows their insignificance to users' satisfaction with the dialogue manager's performance. Hence it follows that the knowledge representation [7], which was incorporated into the dialogue-management process in the second WOZ system, with its rather consistent flexibility in directing the user to select relevant, available data, has no (negative) influence on users' satisfaction.

In addition, we thought that it would be very interesting to see which parameters are significant for individual user-satisfaction metrics. First, we discovered that **Future Use** could not be well modeled in the first experiment and that **User Expertise** and **Interaction Pace** could not be well modeled in the second experiment, the corresponding MLR models explained less than 10% of the variance. Second, the parameters that most significantly predicted the remaining, individual user-satisfaction measures are given in Table 4.

Surprisingly, in the first experiment, the database parameter NND was most significant for all the individual user-satisfaction measures, but, in the second experiment, it was insignificant for all of them. This can be seen as the conclusive evidence of the dialogue manager's cooper-

	WOZ1	WOZ2
TTS Performance	NND ($p < 0.00005$)	UMN ($p < 0.004$)
ASR Performance	NND ($p < 0.00005$)	CMR ($p < 0.012$)
Task Ease	NND ($p < 0.002$)	GDR ($p < 0.02$)
System Response	NND ($p < 0.0003$)	CMR ($p < 0.0002$)
Expected Behaviour	NND ($p < 0.00005$)	Comp, RDR, CMR ($p < 0.04$)

Table 4. Most significant predictors of the individual user-satisfaction measures in the first (WOZ1) and the second (WOZ2) WOZ experiment.

ativity subject to the incorporated knowledge representation [7]. Moreover, all the parameters that were most significant to an individual user-satisfaction measure in the second experiment were insignificant to the same measure in the first experiment. On the one hand, this could indicate that the selected individual user-satisfaction measures really measure the performance of the dialogue manager and consequently illustrate the obvious difference between both dialogue-management manners. On the other hand, one could argue that this simply means that the individual user-satisfaction measures are not appropriate measures of attitude because people are likely to vary in the way they interpret the item wording [9]. Though, due to the huge difference in significance, this seems an unlikely explanation.

5. Conclusion

In this study we have presented the highly portable dialogue-manager component of the developing, bilingual-spoken dialogue system for weather information retrieval. The data gathered in two WOZ experiments was evaluated with the PARADISE framework. This evaluation resulted in several performance equations predicting different dependent variables, all trying to express the performance of the dialogue manager. The observed differences between the derived performance equations make demands upon further empirical research. Not only does a reliable user-satisfaction measure that would capture the performance-measures of different dialogue-system components need to be established, but the reasons for the possible differences between several performance equations also need to be understood and properly assessed.

References

[1] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, A. Anderson, HCRC Dialogue Structure Coding Manual. *Research paper 82, Human Communication Research Centre*, University of Edinburgh, Scotland, (1996).

[2] N. Dahlbäck, A. Jönsson, L. Ahrenberg, Wizard of Oz studies – why and how. *Knowledge-Based Systems* Vol. 6, No. 4 (1993), pp. 258–256.

[3] B. Di Eugenio, M. Glass, The Kappa statistic: a second look. *Computational Linguistics* Vol. 30, No. 1 (2004), pp. 95–101.

[4] A. Ferreira-Cabrera, J. A. Atkinson-Abutridy, A Model for Generating Explanatory Web-based Natural-Language Dialogue Interactions for Document Filtering. *Journal of Research and Practice in Information Technology* Vol. 34, No. 1, (2002), pp. 2–19.

[5] N. M. Fraser, G. N. Gilbert, Simulating Speech Systems. *Computer Speech and Language* Vol. 5, No. 1, (1991), pp. 81–99.

[6] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, S. Busayapongchai, A Form-Based Dialogue Manager for Spoken Language Applications. Presented at the *Proceedings of the International Conference on Spoken Language Processing*, (1996), Philadelphia, USA.

[7] M. Hajdinjak, F. Mihelič, Information-providing dialogue management. In *Lecture Notes in Artificial Intelligence 3206: Text, Speech and Dialogue* (P. Sojka, I. Kopecek, K. Pala, Eds.), Berlin, Springer, (2004), pp. 595–602.

[8] M. Hajdinjak, F. Mihelič, Conducting the Wizard-of-Oz experiment. *Informatica* Vol. 28, No. 4, (2004), pp. 425–429.

[9] M. Hajdinjak, F. Mihelič, The PARADISE Evaluation Framework: Issues and Findings. *Computational Linguistics* Vol. 32, No. 2, (2006), pp. 263–272.

[10] K. S. HONE, R. GRAHAM, Towards a tool for the Subjective Assesment of Speech System Interfaces (SASSI). *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems* Vol. 6 No. 3-4 (2000), pp. 287–303.

[11] G. HOUGHTON, S. D. ISARD, Why to speak, what to say and how to say it: Modelling language production in discourse. In *Modelling Cognition* (P. Morris, Ed.), (1987), pp. 249–267. John Wiley & Sons, New York.

[12] C. KAMM, M. WALKER, D. LITMAN, Evaluating spoken language systems. Presented at the *Proceedings of American Voice Input/Output Society*. (1999), San Jose, USA.

[13] J. KOWTKO, S. D. ISARD, Conversational Games Within Dialogue. *Research paper 31, Human Communication Research Centre*, University of Edinburgh, Scotland, (1992).

[14] S. LARSSON, D. TRAUM, Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, No. 6, (2000), pp. 323–340.

[15] D. J. LITMAN, P. SHIMEI, P Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, No. 12, (2002), pp. 111–137.

[16] K. PEPELNJAK, F. MIHELIČ, N. PAVEŠIĆ, Semantic decomposition of sentences in the system supporting flight services. *Journal of Computing and Information Technology*, Vol. 4, No. 1, (1996), pp. 17–24.

[17] M. POESIO, D. R. TRAUM, Conversational actions and discourse situations. *Computational Intelligence*, Vol. 13, No. 3, (1997), pp. 309–349.

[18] R. POWER, The organization of purposeful dialogues. *Linguistics*, No. 17, (1979), pp. 107–152.

[19] G. A. F. SEBER, *Linear Regression Analysis*. John Wiley & Sons, New York, (1977).

[20] B. G. TABACHNICK, S. FIDELL, *Using Multivariate Statistics* (3rd ed.). Harper Collins, New York, (1996).

[21] M. A. WALKER, D. LITMAN, C. A. KAMM, A. ABELLA, PARADISE: A General Framework for Evaluating Spoken Dialogue Agents. Presented at the *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, (1997), Madrid, Spain.

[22] M. A. WALKER, D. J. LITMAN, C. A. KAMM, A. ABELLA, Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, Vol. 12, No. 3, (1998), pp. 317–347.

[23] M. A. WALKER, C. KAMM, D. LITMAN, Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, Vol. 6, No. 3-4, (2000), pp. 363–377.

[24] V. ZUE, S. SENEFF, J. GLASS, J. POLIFRONI, C. PAO, T. J. HAZEN, L. HETHERINGTON, JUPITER: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, (2000), pp. 85–96.

[25] J. ŽIBERT, S. MARTINČIĆ-IPŠIĆ, M. HAJDINJAK, I. IPŠIĆ, F. MIHELIČ, Development of a Bilingual Spoken Dialog System for Weather Information Retrieval. Presented at the *Procedings of the 8th European Conference on Speech Communication and Technology*, (2003), Geneva, Switzerland.

Received: January, 2006
Accepted: January, 2007

Contact addresses:

Melita Hajdinjak
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 10000 Ljubljana, Slovenia
e-mail: `melita.hajdinjak@fe.uni-lj.si`

France Mihelič
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 10000 Ljubljana, Slovenia

MELITA HAJDINJAK is an Assistant of Mathematics and a Researcher at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. She received her Ph.D. in Electrical Engineering in 2006 and is currently working toward her Ph.D. in Mathematics.

FRANCE MIHELIČ is an Associate Professor at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. His areas of interest include pattern recognition, speech recognition and understanding, speech synthesis and signal processing.