The PARADISE Evaluation Framework: Issues and Findings

Melita Hajdinjak*
University of Ljubljana

France Mihelič*
University of Ljubljana

There has been a great deal of interest over the past 20 years in developing metrics and frameworks for evaluating and comparing the performance of spoken-language dialogue systems. One of the results of this interest is a potential general methodology, known as the PARADISE framework. This squib highlights some important issues concerning the application of PARADISE that have, up to now, not been sufficiently emphasized or have even been neglected by the dialogue-system community. These include considerations regarding the selection of appropriate regression parameters, normalization effects on the accuracy of the prediction, the influence of speech-recognition errors on the performance function, and the selection of an appropriate user-satisfaction measure. In addition, it gives the results of an evaluation of data from two Wizard-of-Oz experiments. These evaluations include different dependent variables and examination of individual user-satisfaction measures.

1. Introduction

A long list of **objective dialogue metrics** (Danieli and Gerbino 1995; Smith and Gordon 1997) for dialogue evaluation, which can be calculated without recourse to human judgment, and **subjective dialogue metrics** (Shriberg, Wade, and Price 1992; Danieli and Gerbino 1995), which are based on human judgments, have been proposed. Their well-known limitations led Walker, Litman, Kamm, and Abella (1997) to propose their paradigm for dialogue system evaluation (PARADISE), a potentially general methodology for evaluating spoken-language dialogue systems, the goal of which was to compare and optimize different dialogue managers and task domains independently.

The PARADISE framework maintains that the system's primary objective is to maximize user satisfaction (Shriberg, Wade, and Price 1992), and it derives a combined performance metric for a dialogue system as a weighted linear combination of **task-success measures** and **dialogue costs.** The dialogue costs are of two types: **dialogue-efficiency costs** (e.g., number of utterances, dialogue time), which are measures of the system's efficiency in helping the user to complete the task, and **dialogue-quality costs** (e.g., system-response delay, mean recognition score), which are intended to capture other aspects that can have large effects on the user's perception of the system's performance.

Applying PARADISE to dialogue data requires dialogue corpora to be collected via controlled experiments during which users subjectively rate their satisfaction. Here, user satisfaction is calculated with a survey (Walker et al. 1998) that asks users to specify the degree to which they agree with several statements about the performance

* Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia

of the system. 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 be hand-labeled. The PARADISE model of performance 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.

The squib addresses some PARADISE issues (with most of them arising from the application of MLR) that have, up to now, not been sufficiently emphasized or have even been neglected by the dialogue-system community. Moreover, most of the considerations about these issues are supported by the results of applying PARADISE to the data from two Wizard-of-Oz (WOZ) experiments (Hajdinjak and Mihelič 2004a) carried out during the development of a weather-information-providing, natural-language spoken dialogue system (Žibert et al. 2003). In contrast to previous PARADISE applications, we evaluated data that were collected in the early stages of a dialogue system's design where speech understanding was simulated by a human.

2. PARADISE-Framework Issues

The PARADISE model of performance (Walker et al. 1998) is defined as follows:

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

Here, α is the weight on the kappa coefficient κ, which is 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 costs c_i, and $\mathcal{N}$ is a Z-score normalization function:

$$\mathcal{N}(x) = \frac{x - \overline{x_0}}{\sigma_{x_0}} \tag{2}$$

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. Normalization guarantees that the weights directly indicate the relative contributions to the performance function, which can be used to predict user satisfaction, i.e., the value of the dependent variable.

2.1 Relationships between MLR Variables

Nevertheless, MLR sets some constraints on the relationships between the parameters (Seber 1977). Most obviously, there must exist at least an approximately linear relationship between the dependent and the independent variables. On the other hand, it can be shown (Patel and Bruce 1995) that dropping the predictors, i.e., the independent variables, that are not statistically significant to the dependent variable (i.e., the p value is greater than 0.05) can reduce the average error of the predictions.

Furthermore, the predictors must not correlate highly (i.e., the absolute values of the correlation coefficients must not be greater than 0.7). If they do, small errors or variations in the values of sample observations can have a large effect on the weights in the performance function (1). Therefore, the redundant predictors must be removed before applying MLR. Obviously, it is reasonable to remove those predictors that are

less statistically significant (i.e., with greater p values) to the dependent variable. Note, high correlations may exist even between variables that seem unrelated (Section 3.2).

2.2 Normalization Effects

MLR is based on the least squares method, where the model minimizes the sum of the squares of the differences between the observed and the predicted values. Hence,

$$\mathcal{N}(\mathrm{US}) = \widehat{\mathcal{N}(\mathrm{US})} + \epsilon \tag{3}$$

where $\mathcal{N}(\mathrm{US})$ is the normalized observed user-satisfaction value, $\widehat{\mathcal{N}(\mathrm{US})}$ is the predicted normalized user-satisfaction value, and ϵ is the error of the prediction. Further,

$$US = \widehat{\mathcal{N}(\mathrm{US})}\sigma_{\mathrm{US}_0} + \overline{\mathrm{US}_0} + \epsilon\sigma_{\mathrm{US}_0} = \widehat{\mathrm{US}} + \epsilon\sigma_{\mathrm{US}_0} \tag{4}$$

is the MLR model for the unnormalized user-satisfaction values, where $\overline{\mathrm{US}_0}$ and σ_{US_0} are the mean value and the standard deviation of the sample set of observed user-satisfaction values, respectively. Note, the initial noise variable ϵ is multiplied by σ_{US_0}.

There are several different measures of goodness of fit for a regression model, but the most widely used is the *coefficient of determination* R^2. In a regression model with normalized variables, R^2 turns out to be the variance of the predicted variable:

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{\mathcal{N}(\mathbf{US}_i)} - \overline{\mathcal{N}(\mathbf{US})})^2}{\sum_{i=1}^{n}(\mathcal{N}(\mathbf{US}_i) - \overline{\mathcal{N}(\mathbf{US})})^2} = \frac{\sum_{i=1}^{n}\widehat{\mathcal{N}(\mathbf{US}_i)}^2}{n} = \mathrm{var}(\widehat{\mathcal{N}(\mathbf{US})}) \tag{5}$$

where $\mathbf{US}_i$ and $\widehat{\mathcal{N}(\mathbf{US}_i)}$ are the ith component of the vector $\mathbf{US}$ of the observed values and the ith component of the vector $\widehat{\mathcal{N}(\mathbf{US})}$ of the predicted normalized values, respectively. One noteworthy consequence of the equality (5) is that the absolute values of the weights in the performance function are not greater than 1.

Nevertheless, the accuracy of the prediction $\widehat{\mathrm{US}}$ for US is indicated by the q ratio:

$$q(\mathrm{US}, \widehat{\mathrm{US}}) = \frac{|\mathrm{US} - \widehat{\mathrm{US}}|}{|\mathrm{US}|} \tag{6}$$

The definitions (2) and (6) and the equality (4) lead to the equality

$$\frac{q(\mathcal{N}(\mathrm{US}), \widehat{\mathcal{N}(\mathrm{US})})}{q(\mathrm{US}, \widehat{\mathrm{US}})} = \frac{\frac{|\mathcal{N}(\mathrm{US}) - \widehat{\mathcal{N}(\mathrm{US})}|}{|\mathcal{N}(\mathrm{US})|}}{\frac{|\mathrm{US} - \widehat{\mathrm{US}}|}{|\mathrm{US}|}} = \frac{\frac{|\mathrm{US} - \overline{\mathrm{US}_0} - \widehat{\mathcal{N}(\mathrm{US})}\sigma_{\mathrm{US}_0}|}{|\mathrm{US} - \overline{\mathrm{US}_0}|}}{\frac{|\mathrm{US} - \widehat{\mathcal{N}(\mathrm{US})}\sigma_{\mathrm{US}_0} - \overline{\mathrm{US}_0}|}{|\mathrm{US}|}} = \frac{|\mathrm{US}|}{|\mathrm{US} - \overline{\mathrm{US}_0}|} \tag{7}$$

which shows that as soon as $\mathrm{US} > \frac{\overline{\mathrm{US}_0}}{2}$ the prediction for the normalized values is (usually by a large margin) not as good as the prediction for the unnormalized values. Therefore, after predicting the normalized user-satisfaction values, these values should be transformed back to the original scale to guarantee more accurate predictions.

In previous work (Walker et al. 1997, 1998; Walker, Kamm, and Litman 2000; Litman and Shimei 2002), not only was there no attention paid to these details, it was not

mentioned that the observed user-satisfaction values need to be normalized as well if one wants an acceptable error in the prediction.

2.3 Choosing the Best Set of Predictors

A major problem in regression analysis is that of deciding which predictors should be in the model. There are two conflicting criteria. First, the model chosen should include as many predictors as possible if reliable predictions are to be obtained (also, R^2 increases with the number of predictors). Second, because of the costs involved in determining a large number of predictors and the benefit of focusing only on the most significant predictors, we would like the equation to include as few predictors as possible.

An appropriate method of choosing the best set of predictors is *backward elimination* (Seber 1977). Here, at each step, a single predictor is eliminated from the current regression model if its removal would increase the sum-of-squares differences between the observed and the predicted values,

$$\mathrm{RSS} = \sum_{i=1}^{n} (\mathcal{N}(\mathbf{US}_i) - \widehat{\mathcal{N}(\mathbf{US}_i)})^2 = \sum_{i=1}^{n} {\epsilon_i}^2 \tag{8}$$

by not more than F_{out} (some properly chosen constant, typically between 2 and 4) times the residual mean square $\frac{\mathrm{RSS}}{n-p}$, where n is the number of observations and p is the number of predictors. That is, the predictor giving the smallest increase in RSS is chosen.

2.4 Why Model the Sum of User-Satisfaction Scores

Hone and Graham (2000) argued that the items chosen for the user-satisfaction survey (Table 1) introduced within the PARADISE framework were based neither on theory nor on well-conducted empirical research. Moreover, they said that the way that the collected data was used would be inappropriate, i.e., the approach of summing all the scores could only be justified on the basis of evidence that all of the items are measuring the same construct, otherwise the overall score would be meaningless.

Table 1
The user-satisfaction survey used within the PARADISE framework.

1. *Was the system easy to understand?* (**TTS Performance**)
2. *Did the system understand what you said?* (**ASR Performance**)
3. *Was it easy to find the message you wanted?* (**Task Ease**)
4. *Was the pace of interaction with the system appropriate?* (**Interaction Pace**)
5. *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?* (**Expected Behavior**)
8. *From your current experience with using the system, do you think you'd use the system regularly when you are away from your desk?* (**Future Use**)

If, in spite of Hone and Graham's remarks, one wants to evaluate user satisfaction with the survey given in Table 1, the question that arises is whether the target to be predicted should really be the sum of all the user-satisfaction scores. However, our experiments (Section 2.3) showed a remarkable, but expected, difference in the significance of the predictors when taking different satisfaction-measure sums or even individual scores as the target to be predicted. Moreover, some individual user-satisfaction measures could not be well modeled. Consequently, we think that it would be more appropriate to take the sum of the user-satisfaction scores that are likely to measure the selected aspect (e.g., dialogue-manager performance) of the system's performance.

2.5 Speech-recognition Effects

It has often been reported (Walker et al. 1998; Kamm, Walker, and Litman 1999; Walker, Kamm, and Litman 2000; Litman and Shimei 2002) that the mean concept accuracy, often referred to as the mean recognition score, is the exceptional predictor of a dialogue system's performance. Moreover, it has been shown that as recognizer performance improves the significance of the predictors can change (Walker, Borland, and Kamm 1999).

Thus, we would go so far as to claim that the influence of automatic speech recognition hinders the other predictors from showing significance when evaluating the performance of the dialogue-manager component, excluding the efficiency of its clarification strategies. Only if the users are disencumbered from speech-recognition errors are they able to reliably assess the mutual influence of the observable less significant contributors to their satisfaction with the dialogue manager's performance. Therefore, in our WOZ experiments (Section 2), which were carried out in order to evaluate the performance of the dialogue manager, speech understanding (i.e., speech recognition and natural-language understanding) was performed by a human. As expected, κ with mean values near 1 (0.94 and 0.98, respectively), which was the only predictor reflecting speech-recognition performance (Section 2.1), did not show the usual degree of significance in predicting the performance of our WOZ systems (Section 2.3).

3. PARADISE Framework Application

With the intention of involving the user in all the stages of the design of the spoken natural language, weather-information-providing dialogue system (Žibert et al. 2003), WOZ data were collected and evaluated even before the completion of all the system's components. The aim of the first two WOZ experiments (Hajdinjak and Mihelič 2004a) was to evaluate the performance of the dialogue-manager component (Hajdinjak and Mihelič 2004b). Therefore, while the task of the wizard in the first WOZ experiment was to perform speech understanding and dialogue management, the task in the second WOZ experiment was to perform only speech understanding, and the dialogue-management task was assigned to the newly implemented dialogue manager.

There were 76 and 68 users involved in the first and the second WOZ experiment, respectively. The users were given two tasks: The first task was to obtain a particular piece of weather-forecast information, and the second task was a given situation, the

aim of which was to stimulate them to ask context-specific questions. In addition, the users were given the freedom to ask extra questions. User satisfaction was then evaluated with the user-satisfaction survey given in Table 1, and a comprehensive **user satisfaction** (US) was computed by summing each question's score and thus ranged in value from a low of 8 to a high of 40.

3.1 Selection of Regression Parameters

The selection of regression parameters is crucial for the quality of the performance equation, and it is usually the result of thorough considerations made during several successive regression analyses. However, following the recommended Cohen's method (Di Eugenio and Glass 2004), we first computed the task-success measure

> **Kappa coefficient** (κ), reflecting the wizard's typing errors and unauthorized corrections,

and the dialogue-efficiency costs

> **Mean elapsed time** (MET), i.e., the mean elapsed time for the completion of the tasks that occurred within the interaction, and
>
> **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.

Second, the following dialogue-quality costs were selected:

> **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 in each of the user's turns;
>
> **Mean response time** (MRT), i.e., the mean system-response time;
>
> **Number of missing responses** (NMR), i.e., the difference between the number of turns by the system and the number of turns by the user;
>
> **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 unexpected responses from the system, including pardon moves;
>
> **Number of errors** (Error), i.e., the number of system errors, e.g., interruptions of the telephone connection and unsuitable natural-language sentences;

Number of help messages (NHM) and **help-message ratio** (HMR), i.e., the number and the ratio of system's help messages;

Number of check moves (NCM) and **check-move ratio** (CMR), i.e., the number and the ratio of system's moves checking some information regarding past dialogue events;

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;

Number of abandoned requests (NAR) and **abandoned-request ratio** (ARR), i.e., the number and the ratio of the information-providing games abandoned by the user.

Note that 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 will refer to them as **database parameters**. It has, however, been argued (Walker et al. 1998) that the database size might be a relevant predictor of performance. However, the decision to introduce the database parameters as dialogue costs was based on the extremely sparse and dynamical weather-information source (Hajdinjak and Mihelič 2004b) with a time-dependent data structure that we had at our disposal.

3.2 Correlations between MLR Parameters

In the data from both WOZ experiments, some high correlations between the regression parameters were observed. Note that the high correlations were not found only between the newly introduced or the obviously related parameters such as NUT and NGD. In the first experiment, not entirely evident high correlations were observed between MET and MRT (0.7), NMR and NHM (0.7), GDR and NDR (-0.8), but in the second experiment between MET and MRT (0.8), NMR and NHM (0.7), GDR and RDR (-0.7). The regression parameters GDR and NDR as well as GDR and RDR were highly correlated only in one of the WOZ experiments, and moderately in the other.

3.3 Performance-function Results

We applied PARADISE to the data from both WOZ experiments (Hajdinjak and Mihelič 2004a). 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) and could, in addition, be well modeled, i.e., the sum of the user-satisfaction-survey scores assigned to ASR Performance, Task Ease, System Response, and Expected Behavior (Table 1).

After removing about 10% of the outliers,[1] backward elimination for $F_{\text{out}} = 2$ was performed. Thus, the data from the first WOZ experiment gave the following performance equations:

$$\begin{aligned}\widehat{\mathcal{N}(\text{US})} &= -0.69\mathcal{N}(\text{NND}) - 0.16\mathcal{N}(\text{NRD}) \\ \widehat{\mathcal{N}(\text{DM})} &= -0.61\mathcal{N}(\text{NND}) - 0.16\mathcal{N}(\text{NRD}) + 0.21\mathcal{N}(\text{Comp})\end{aligned}$$

with 58% ($R^2 = 0.58$) and 59% of the variance explained, respectively. To be able to see the difference between these two equations, note that Comp was significant for US ($p < 0.02$), but removed by backward elimination. In contrast, the data from the second WOZ experiment gave the following performance equations:

$$\begin{aligned}\widehat{\mathcal{N}(\text{US})} &= -0.30\mathcal{N}(\text{CMR}) + 0.18\mathcal{N}(\kappa) - 0.23\mathcal{N}(\text{MET}) \\ \widehat{\mathcal{N}(\text{DM})} &= -0.35\mathcal{N}(\text{CMR}) + 0.35\mathcal{N}(\kappa) + 0.35\mathcal{N}(\text{GDR}) - 0.17\mathcal{N}(\text{ARR})\end{aligned}$$

with 26% and 46% of the variance explained, respectively. Note, MET was significant for DM ($p < 0.02$) and that GDR and ARR were significant for US ($p < 0.04$), but they were all removed by backward elimination. Moreover, MET was trivially correlated with GDR and ARR (i.e., the correlation coefficients were lower than 0.1).

Let us compare both performance equations predicting DM. The first observation 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, NRD, Comp) 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. Hence it follows that the developed dialogue manager (Hajdinjak and Mihelič 2004b) with its rather consistent flexibility in directing the user to select relevant, available data does not (negatively) influence users' satisfaction.

Moreover, we thought that it would be very interesting to see which parameters are significant for individual user-satisfaction measures (Table 1). However, the situation in which the Task Ease question was the only measure of user satisfaction, the aim of which was to maximize the relationship between elapsed time and user satisfaction, was considered before (Walker, Borland, and Kamm 1999). First, we discovered that Future Use could not be well modeled in the first WOZ experiment and that User Expertise and Interaction Pace could not be well modeled in the second WOZ experiment, i.e., the corresponding MLR models explained less than 10% of the variance.

Second, the parameters that most significantly contributed to the remaining, individual user-satisfaction measures are given in Table 2. Surprisingly, the parameters that were most significant to an individual user-satisfaction measure in the first WOZ

1 Eliminating outliers, i.e., observations that lie at an abnormal distance from other values, is a common practice in multivariate linear regression (Tabachnick and Fidell 1996). However, before the elimination of the outliers in the data from the first WOZ experiment, the MLR models explained 44% and 39% of the variance, respectively. In contrast, before the elimination of the outliers in the data from the second WOZ experiment, the MLR models explained 18% and 34% of the variance, respectively.

Table 2
Most significant predictors of the individual user-satisfaction measures in the first (*WOZ1*) and the second (*WOZ2*) WOZ experiment.

	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 Behavior	NND ($p < 0.00005$)	Comp, RDR, CMR ($p < 0.04$)

experiment were insignificant to the same measure in the second WOZ experiment and vice versa. 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 (Hone and Graham 2000). However, due to the huge difference in significance the latter seems an unlikely explanation.

4. Conclusion

The application of PARADISE to the data from two WOZ experiments led us to the following conclusions. First, the identified high correlations between some dialogue costs and the explained normalization effects on the accuracy of the prediction reinforce the need for careful regression analysis. Second, if speech recognition is performed by a human, the PARADISE evaluation will lead to the identification of the significant predictors of the dialogue-manager's performance. Third, inaccurate MLR models of some individual user-satisfaction measures and the observed differences between pairs of performance equations predicting the same dependent variable require 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

Danieli, Morena and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, Stanford.

Di Eugenio, Barbara and Michael Glass. 2004. The Kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Hajdinjak, Melita and France Mihelič. 2004a. Conducting the Wizard-of-Oz experiment. *Informatica*, 28(4): 425–430.

Hajdinjak, Melita and France Mihelič. 2004b. Information-providing dialogue management. In *Proceedings of TSD*, pages 595–602, Brno.

Hone, Kate S. and Robert Graham. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 6(3–4):287–303.

Kamm, Candace A., Marilyn A. Walker, and Diane J. Litman. 1999. Evaluating spoken language systems. In *Proceedings of American Voice Input/Output Society*, San Jose.

Litman, Diane J. and Pan Shimei. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12:111–137.

Patel, Nitin R. and Peter C. Bruce. 1995. *Resampling Stats—Data Mining in Excel: Lecture Notes and Cases*. Trafford Holdings Ltd., Victoria, Canada.

Seber, George A. F. 1977. *Linear Regression Analysis*. John Wiley & Sons, New York.

Shriberg, Elizabeth, Elizabeth Wade, and Patti Price. 1992. Human-machine problem solving using Spoken Language Systems (SLS): Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and NL Workshop*, pages 49–54, New York.

Smith, Ronnie W. and Steven A. Gordon. 1997. Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialog. *Computational Linguistics*, 23(1):141–168.

Tabachnick, Barbara G. and Linda S. Fidell. 1996. *Using Multivariate Statistics*, 3rd ed. Harper Collins, New York.

Walker, Marilyn A., Julie Borland, and Candace A. Kamm. 1999. The utility of elapsed time as a usability metric for spoken dialogue systems. In *Proceedings of ASRU*, pages 317–320, Keystone.

Walker, Marilyn A., Candace A. Kamm, and Diane J. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 6(3–4):363–377.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of ACL*, pages 271–280, Madrid.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12(3):317–347.

Žibert, Janez, Sanda Martinčić-Ipšić, Melita Hajdinjak, Ivo Ipšić, and France Mihelič. 2003. Development of a bilingual spoken dialog system for weather information retrieval. In *Proceedings of EUROSPEECH*, pages 1917–1920, Geneva.