End-to-End Evaluation of Multimodal Dialogue Systems - can we Transfer Established Methods?

Nicole Beringer[1], **Katerina Louka**[1], **Victoria Penide-Lopez**[2], **Uli Türk**[1]

[1] Institut für Phonetik und Sprachliche Kommunikation
Schellingstr. 3, D-80799 München, Germany
[2] BIOMAX Informatics AG.
Lochhamer Str. 11, D-82152 Martinsried, Germany
{beringer,kalo,penide,tuerk}@phonetik.uni-muenchen.de

Abstract

The goal of this paper is to define a methodology for the end-to-end evaluation of the multimodal dialogue system SmartKom along the lines of the DARPA guidelines for spoken dialogue systems. The methodology consists of an extended framework for the evaluation of a multimodal dialogue system, evaluation metrics for its various components, and an approach to compare the user satisfaction with the system's technical performance.

1. Introduction

The paper gives an outline of how to generalize task requirements from spoken to multimodal dialogues and it offers a possibility to compare dialogue strategies as well as to normalize performance for task complexity of different tasks. It also reports on the problems we had with transfering the PARADISE (Walker et al., 1997) framework to our multimodal system and how we resolved them.

The following section describes shortly the function of the multimodal SmartKom dialogue system which has to be evaluated. Section 3 gives a general outline on end-to-end evaluation requirements. In section 4 we describe which task requirements can be used with regard to multimodality vs. unimodality and across tasks. Special attention is turned to the application of PARADISE on the SmartKom system. The most challenging problems due to multimodality are described in section 5. Section 6 reports our approach to solve the evaluation problems due to multimodality: PROMISE. The last section gives an outlook on PROMISE.

2. The SmartKom project

In the SmartKom project an intelligent computer-user interface is being developed which deals with various kinds of oral or physical input. Potential benefits of SmartKom include the ease of use and the naturalness of the man-machine interaction due to multimodal input and output. However, a very critical obstacle to progress in this area is the lack of a general methodology for evaluating and comparing the performance of the three possible scenarios provided by SmartKom:

- SmartKom Home/Office to communicate and operate machines at home (e.g. TV, workstation, radio),
- SmartKom Public to have a public access to public services, and
- SmartKom Mobile as a mobile assistant.

The system understands input in the form of natural speech as well as in the form of gestures. In order to ”react” properly to the intentions of the user, the emotional status is analyzed via the facial expression and the prosody of speech. One of the requirements of the project is to develop new modalities and new techniques.

3. General Outline on End-to-End-Evaluation

Because of the innovative character of the project, new methods for end-to-end evaluation had to be developed partly through transferring established criteria from the evaluation of spoken dialogue systems, and partly through the definition of new multimodal measures. These criteria have to deal with a fundamental property of multimodal dialogue systems, namely the high variability of the input and output modalities with which the system has to cope.

The performance of the evaluation is very often driven by the characteristics of the system that has to be judged (Andenfilger, 1997). For the SmartKom Evaluation we have to take three aspects into account:

- the needs of the developers,
- users' needs and
- the constraints on the evaluation of multimodal systems in general.

We tried to combine these three aspects in our concept as well as in the performance of the evaluation by computing the efficiency of the system on the basis of objectively measurable criteria such as duration of the dialogue, as well as on the basis of subjective criteria, such as user satisfaction and acceptance by the user as follows:

3.1. The developers' needs

An end-to-end evaluation must focus on the quality of man-machine-interaction. The main goal thereby is to deliver reliable results of the performance of the multimodal system in question, under realistic application conditions, in order to enable an improvement of the system as a whole or parts of it. Due to multimodality, we have to pay special attention to the correlation of the users' input modalities and the output of the system.

3.2. The users' needs

The aim of the SmartKom system is to provide different services to the user, which can be retrieved multimodally and quasi naturally. Via questionnaire, the users' needs are investigated and evaluated according to user satisfaction.

3.3. Constraints on the evaluation of multimodal systems in general

Although our project aims to offer more natural man-machine interactions there are some constraints which we have to consider while evaluating the system. We have to distinguish between naturalness of input and naturalness of output. Technical progress in ASR, gesture recognition and recognition of facial expression improved in recent years, so did the technical realisation of display or speech synthesis. Nevertheless, combining the latter, the developers have to create an agent as a whole which is both competent, virtual and multimodal!

Of course, multimodal dialogue system evaluation has to cope with the following standard problems of dialogue evaluation:

- How can we abstract from the system itself, i.e. the different hardware and software components, in order to evaluate across dialogues and scenarios (see above?
- How can we abstract from different dialogue strategies?

In section 6 we will present our solution to these problems.

3.4. Feedback during Evaluation - Preliminary Evaluation

For every scenario developers get access to all content-related problems, results and protocols within the correspondent evaluation phase. For all results we offer a comparison of usability evaluation and objectively measured values via a graphical evaluation tool (Beringer et al., 2002a). This allows the developers to precisely follow the most important results to obtain the maximal improvement.

General problems of the system are also made public to the developers. This makes it possible to find the weak points of the system, to improve the performance of the involved modules or module clusters, and to install updates whenever possible throughout this evaluation phase.

3.5. Feedback after Evaluation - Final Evaluation

For each scenario we will evaluate the system under the same circumstances in a final evaluation. For the defined evaluation period the system cannot be optimized for evaluation. Having finished one evaluation period, the results are published to the project members.

4. Task Requirements

In order to evaluate a multimodal system -in particular SmartKom- we had to extend a unimodal dialogue evaluation to adapt it to the requirements of multimodality. As a result, a number of components did not fit into a monomodal dialogue evaluation like PARADISE and we had several problems in defining reference material. The PARADISE framework gives a useful and promising approach to comparing different dialogue strategies and different spoken dialogue systems via attribute value matrices (AVM), to compute the task sucess measure, to define several quality and quantity measures socalled cost functions, and to weigh their importance for the perfomance of the system via multiple linear regression, dependent on the User Satisfaction value.

In SmartKom the user is given a rather unprecise task definition, in order to enable an interaction between the user and the system which is as natural as possible. The widely used approach of providing a set of reference keys (Hirschmann et al., 1990) was not possible in our case.

For example in the case of an electronic programming guide - one of the SmartKom tasks - we would get a sparse two-dimensional matrix with up to several thousand rows/columns, in which each of the values is potentially correct. The possibility of a mismatch between actual value and key is much higher, and, considering the unprecisely defined task description, is possibly not in the range of ”error”. Our solution is to extract different superordinate concepts depending on the task at hand and to compute a bipolar function in order to handle the possible reference answers which otherwise would be made diffuse.

For example, when planning an evening watching TV, these superordinate concepts, which we termed ”information bits”, would contain movie title, genre, channel, timeslots, actors etc. Similar to a content-analysis, these ”information bits” are carefully selected, categorized and weighted by hand before the tests start. This enables us to compute, normalize and compare across different tasks and scenarios.

Using this approach, we are independent of static reference keys. In conjunction with the extended attribute value matrix it is not quite clear how to obtain an optimal dialogue or task length.

5. Problems due to Multimodality

The advance of the SmartKom system, namely multimodality, turned out to be a big problem for the end-to-end evaluation of the system. Due to the multimodal input and output facilities, a man-machine dialogue becomes very complex. It is not quite clear how to evaluate the recognition and interpretation of facial expression, gestural input, prosodic cues or the efficiency of the multimodal interaction according to the user's needs and preferences.

We had to find a possibility to define the dependencies of the different modalities among each other and weigh them accordingly. Due to the different complexities of recognition and interpretation of speech, gesture and facial expression we also had to prove that none of the modalities was in its use.

5.1. How to score multimodal inputs or outputs?

In contrast to interactive unimodal spoken dialogue systems, which are based on many component technologies like speech recognition, text-to-speech, natural language understanding, natural language generation and database query languages, multimodal dialogue systems consist of several such technologies which are functionally similar to each other and therefore could interfere with each other. To make this clear, just imagine the similar functions of ASR and Gesture Recognition: while interacting with a multimodal man-machine interactive system like SmartKom users have the posibility to say what information they want to have and to simultaneously give the same, an additional, or a more specific input via ”interactional gesture” (Steininger et al., 2001). There are several open questions to multimodal evaluation:

- For the purpose of evaluation, are the multimodal inputs considered synchronous or are they timed different?
- Are inputs from different modalities equivalent, i.e. are they describing the same user intention, although they may not be synchronous in time?
- Does the system have to cope with different inputs?

There are several possible problem solving strategies for the system namely:

- First match: the information which was recognized first is taken for further system processing, regardless of the recognition method. This would of course not help in multimodal processing.
- ”Mean” match: the system takes the information which is common to both of the recognition modules. This could be called multimodal verification.
- Additional match: take all the information given by several recognizers for further system processing. This would be the best solution, if we assume all recognizers to be highly accurate.

As described in section 6 and (Beringer et al., 2002b) we correlated the accuracy of the chosen problem solving strategy in this case with a corresponding question in the survey for evaluation.

5.2. How to weight the several multimodal components of recognition systems?

How can we estimate the accuracy of different recognizers? I.e., in talking about speech recognition, we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizible gestures which can be found in a given coordinate plane. It should be clear, that

- the gesture recognizer, for example, will be more accurate than the ASR system but
- the performance of the ASR system must get a higher value than the one of gesture recognition within the system!

Please note, that the recognition of prosodic information as well as the recognition of facial expression are only used additionally to define the user state in order for the system to react properly. Therefore, these two have to get a lower weight in the calculation.

6. PROMISE - general description of a multimodal framework

As we have shown when describing the problems problems in the preceding sections evaluating multimodal systems cannot mean transferring established methods from spoken dialogue evaluation like the PARADISE framework one-to-one.

To make clear, that multimodality causes difficulties which cannot be handled by defining some additional variables to compute the performance of a multimodal dialogue system and therefore the evaluations of spoken and multimodal dialogue systems cannot be compared in a strict sense, we decided to define a multimodal dialogue evaluation framework PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation). For further details please refer to (Beringer et al., 2002b). This new framework of course uses established methods from spoken dialogue evaluations but has to take into account new methods to handle multimodal characteristics like gestural input combined with speech input, graphical vs. speech output or userstate information via facial expression of the user.

6.1. How to implement Usability?

In order to evaluate the system performance it is necessary to collect information about the operation and the usability of the system. Without the involvement of the user no successful evaluation can be executed. The three prerequisites for the goal of achieving user satisfaction are:

- usability,
- efficiency and
- user’s acceptance.

In order to get the relevant data we make use of the questionnaire technique. After each task, users are required to fill out a questionnaire in web page form. The questionnaire is adjusted to the different scenarios provided by SmartKom. It includes and extends the usability’s survey given by the PARADISE framework (Walker et al., 1997). It is subdivided into three sections:

- In the first section, users are asked to give information about themselves
- The second section includes questions about operating and communicating with the system
- The third section contains inquiries about the future use/users of the system.
- The information needed for the calculation of user satisfaction is mostly extracted from the second section of the questionnaire.

Due to the multimodality of SmartKom, the goal of the questions is to evaluate both the oral and graphical output of the system. The questionnaire consists of three types of questions:

- questions to which users respond in their own words

- multiple choice questions which can be combined with further statements of the user
- multiple choice questions

Most responses to multiple choice questions ranged over values from 3+ to 3-.The scala consists of seven grades, where +++ stands for “I perfectly agree”, ++ is the shortcut for “I agree, but there are small deviations” and so on. +/- marks indecision.

For further processing each of these responses is mapped to an integer in the range between 3 and -3. Open questions are evaluated by hand. This means that objective measurable costs will be adressed in each questionnaire. User satisfaction is correlated to the performance of the system. The problem is to specify how to quantify/weigh these measures in the evaluation of the overall performance.

6.2. How to generate an objective evaluation of multimodal dialogues?

Our solution to generate an objective evaluation of multimodal dialogue systems, is to define quality and quantity measures (further referred to as costs) which can be measured during system processing and to weight them accordingly. Tables 1 and 2 give an overview of the costs we defined for the SmartKom evaluation. Generally, all costs are matched to the corresponding results of a usability questionnaire.

6.3. Solutions

To calculate the performance over the different applications of the system, we had to generalize the performance affecting factors: transaction success, dialogue strategy, task factors like database size or environmental factors such as background noise, inadequate lighting etc. To be able to compare different dialogue situations and system status, this generalization is done by a normalization function :

$$\mathcal{N}(c_i) = \frac{c_i - \bar{c_i}}{\sigma_{c_i}}$$

where c_i is cost i,
$\bar{c_i}$ is the mean of c_i,
σ_{c_i} is variance of c_i.
Because costs are not equally likely, we defined weights which correlate user satisfaction with the corresponding cost via the statistic Pearson correlation function. Recent work refined the PROMISE framework (see (Beringer et al., 2002b)). All these factors build up the performance formula which is

$$\text{performance} = \alpha\bar{\tau} \ - \sum_{i=1}^{n} \omega_i \mathcal{N}(c_i)$$

with α the Pearson correlation between τ_j(task success),
$\bar{\tau}$ the mean value of all τ_j with j index of tests, and the values
$\tau_j = +1$: task success;
$\tau_j = -1$: task failure;
n the maximum number of different cost function indexes,
c_i the assumed Gaussian cost functions - consistently either mean or cumulative sum of one cost category i (measured over all tests)
weighted by ω_i - the Pearson correlation between cost function c_i and defined associated user satisfaction value, and $\mathcal{N}(c_i)$ the z-score normalization function.

In contrast to PARADISE where a successful task can be computed out of an attribute value matrix over a static set of dialogue keys[2], we defined a set of information bits - superordinate concepts of information, which can vary in number. Reducing τ_j to the bipolar function defined above, we are capable to handle the relevant information out of information bits, which are the basis for the completion of a task. For further details please refer to (Beringer et al., 2002b).

For costs that did not correspond with a question of the usability questionnaire, we defined weights quasi-

Quality measures	
system-cooperativity	measure of accepting misleading input
semantics	no.of multiple input possible misunderstandings of input/output semantical correctness of input/output
helps	no. of offered help for the actual interaction situation
recognition	speech facial expression gestures
transaction success	no. of completed sub-tasks
diagnostic error messages	percentage of error prompts
dialogue complexity	task complexity (needed information bits for one task) input complexity (used information bits) dialogue manager complexity (presentation of results)
ways of interaction	gestures/graphics vs.speech n-way communication (several modalities possible at the same time?)
synchrony	graphical and speech output
user/system turns	mixed initiative dialogue management incremental compatibility

Table 1: Quality measures for the SmartKom evaluation

[1] (Oppermann et al., 2001)

[2] A key defines a necessary information for completing a task successfully

Quantity measures

barge-in	no. of user and system overlap by means of backchanelling, negation of output, further information
cancels	planned system interrupts due to barge-in
off-talk[1]	no. of non-system directed user utterances
elapsed time	duration of input of the facial expression duration of gestural input duration of speech input duration of ASR duration of gesture recognition mean system response time mean user response time task completion duration of the dialogue
rejections	error frequency of input which require a repetition by the user
timeout	error rate of output error rate of input
user/system turns	no. of turns no. of spoken words no. of produced gestures percentage of appropriate/ inappropriate system directive diagnostic utterances percentage of explicit recovery answers

Table 2: Quantity measures for the SmartKom evaluation

objectively via the graphical evaluation tool presented in figure 1 (for details please refer to (Beringer et al., 2002a)).

An evaluator defines the force of the weight by comparing real data and cost information. This supplementing tool allows to give a precise feedback to the developers in the preliminary evaluation phase. With this tool the developers can easily find out which evaluation part (user satisfaction of objective evaluation) to follow in each situation.

7. Conclusion and Future Work

One objective was to provide a tool for judging the quality of the SmartKom system according to objective and subjective efficiency in order to obtain the optimal balance between the needs of the developers and those of the users. Therefore, we analysed the subjective user satisfaction components and compared them with the real behaviour of the system. The output is a normalized balance between subjective and objective criteria.

Another objective was to design, build and test a set of tools that allow us to easily compare system properties and evaluate them according to PROMISE (Beringer et al., 2002b) and PARADISE criteria (Walker et al., 1997). Therefore, we developed tools to analyse different potential dialogue strategies for carrying out a task like finding a hotel, planning a guided tour or compiling a personal TV programme. We had to provide measures such as inappropriate utterance ratio, turn correction ratio, concept accuracy, implicit recovery and transaction success, performance over subdialogues and dialogues or normalization of the performance for task complexity along the lines of the PROMISE framework.

We defined an adequate evaluation strategy for multimodal systems, taking into account established methods from spoken dialogue evaluations like PROMISE as well as methods to get along with the existing transfer problems of mono- vs. multimodality.

8. Acknowledgements

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the SmartKom project (01IL905E/6).

9. References

U. Andenfilger. 1997. Some remarks about the validation of information systems development. in interdisciplinary foundation of systems design and evaluation. Seminar report, Center for Computer Science, edited by R. Keil-Slawik, Saarbrücken Schloss Dagstuhl, University of Saarland.

N. Beringer, S.Hans, K. Louka, and J. Tang. 2002a. How to relate user satisfaction and system performance in multimodal dialogue situations? - a graphical approach.

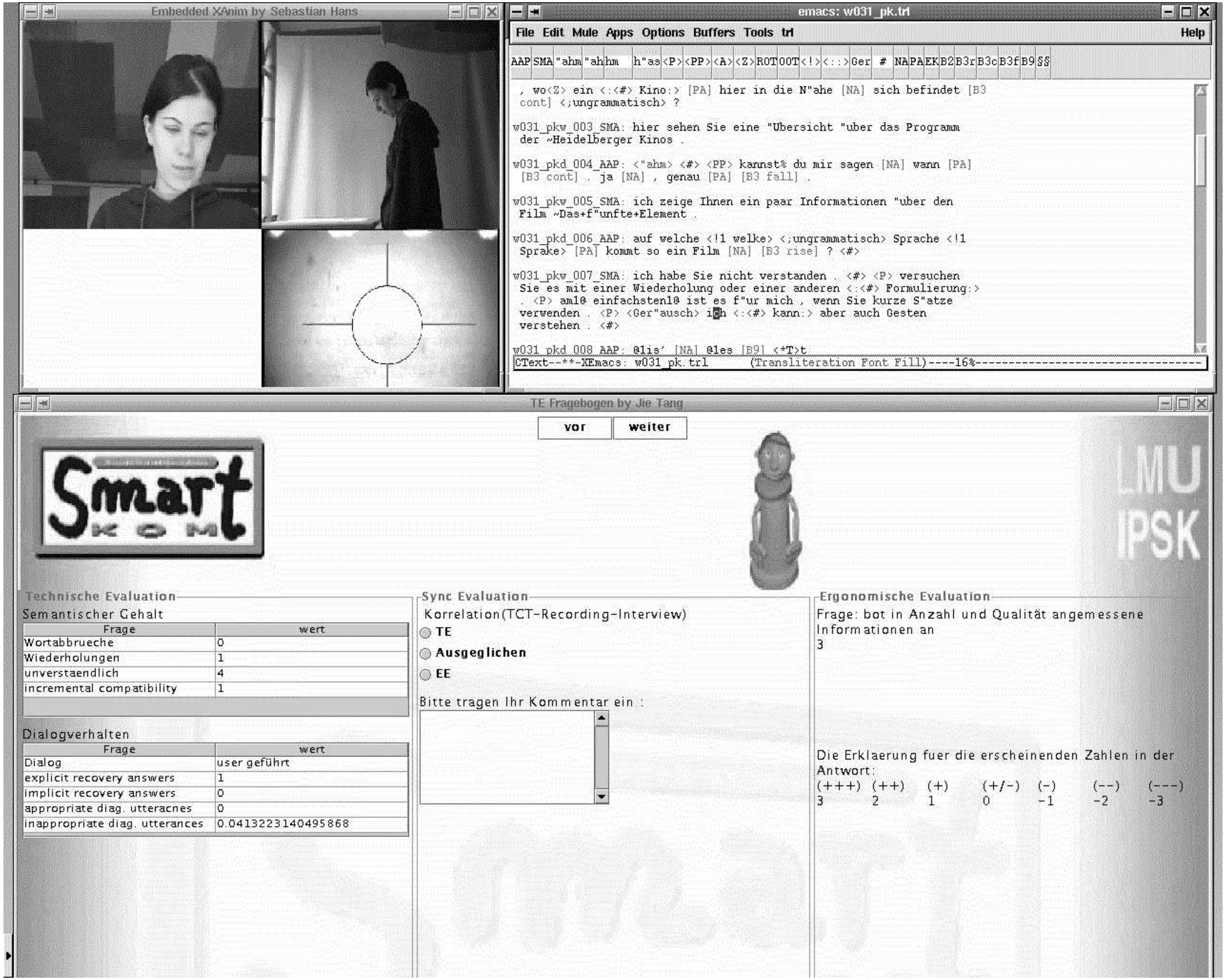
Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, and Uli Türk. 2002b. Promise - a procedure for multimodal interactive system evaluation. *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation (to appear)*.

L. Hirschmann, D.A. Dahl, D.P. McKay, L.M. Norton, and M.C. Linebarger. 1990. Beyond class a: A proposal for automatic evaluation of discourse. *Proceedings of the Speech and Natural Language Workshop*.

D. Oppermann, F. Schiel, S. Steininger, and N. Beringer. 2001. Off-talk - a problem for human-machine-interaction. *Proc. of EUROSPEECH 2001, Scandinavia, Aalborg*.

S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in smartkom - the coding system. *Springer Gesture Workshop 2001, London (to appear)*, LNAI 2298.

M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluation spoken dialogue agents. *Annnual Meeting of the Association of Computational Linguistics. ACL.*



Embedded XAnim by Sebastian Hans
emacs: w031_pk.trl
File Edit Mule Apps Options Buffers Tools trl Help
AAP SMA "ahm "ah hm h"as <P> <PP> <A> <Z> ROT OOT <!> <::> Ger # NA PA EK B2 B3r B3c B3f B9 §§
TE Fragebogen by Jie Tang
vor
weiter
Smart KOM
LMU
IPSK

```
 , wo<Z> ein <:<#> Kino:> [PA] hier in die N"ahe [NA] sich befindet [B3
 cont] <;ungrammatisch> ?

w031_pkw_003_SMA: hier sehen Sie eine "Ubersicht "uber das Programm
 der ~Heidelberger Kinos .

w031_pkd_004_AAP: <"ahm> <#> <PP> kannst% du mir sagen [NA] wann [PA]
 [B3 cont] . ja [NA] , genau [PA] [B3 fall] .

w031_pkw_005_SMA: ich zeige Ihnen ein paar Informationen "uber den
 Film ~Das+f"unfte+Element .

w031_pkd_006_AAP: auf welche <!1 welke> <;ungrammatisch> Sprache <!1
 Sprake> [PA] kommt so ein Film [NA] [B3 rise] ? <#>

w031_pkw_007_SMA: ich habe Sie nicht verstanden . <#> <P> versuchen
 Sie es mit einer Wiederholung oder einer anderen <:<#> Formulierung:>
 . <P> am1@ einfachsten1@ ist es f"ur mich , wenn Sie kurze S"atze
 verwenden . <P> <Ger"ausch> ich <:<#> kann:> aber auch Gesten
 verstehen . <#>

w031_pkd_008_AAP: @1is' [NA] @1es [B9] <*T>t
CText--**-XEmacs: w031_pk.trl      (Transliteration Font Fill)----16%-----------------------------------
```

Technische Evaluation

Semantischer Gehalt

Frage	wert
Wortabbrueche	0
Wiederholungen	1
unverstaendlich	4
incremental compatibility	1

Dialogverhalten

Frage	wert
Dialog	user geführt
explicit recovery answers	1
implicit recovery answers	0
appropriate diag. utteracnes	0
inappropriate diag. utterances	0.0413223140495868

Sync Evaluation

Korrelation(TCT-Recording-Interview)

○ TE

○ Ausgeglichen

○ EE

Bitte tragen Ihr Kommentar ein :

Ergonomische Evaluation

Frage: bot in Anzahl und Qualität angemessene Informationen an

3

Die Erklaerung fuer die erscheinenden Zahlen in der Antwort:

(+++)	(++)	(+)	(+/-)	(-)	(--)	(---)
3	2	1	0	-1	-2	-3

Figure 1: Graphical evaluation tool for the SmartKom evaluation