A XML-based tool for evaluation of SLDS

Marcela Charfuelán[1], Luis Hernández Gómez, Cristina Esteban López, Holmer Hemsen[1]

Signals Processing Applications Group (GAPS)
Universidad Politécnica de Madrid, Dep. SSR-ETSIT
Ciudad Universitaria s/n 28040, Madrid, Spain
{luis, cristina}@gaps.ssr.upm.es
{marcela, hemsen}@nis.sdu.dk

Abstract

This paper addresses two topics relevant to the evaluation of Spoken Language Dialogue Systems (SLDSs): methodology and tools. We present a methodology for evaluation of SLDSs which includes formalising of procedures for annotation, representation and processing of spoken dialogues for evaluation. Also we present a tool with which to carry on most of the procedures usually applied in evaluation of SLDS nowadays.

1. Introduction

Our experience on evaluation of spoken language dialogue systems indicates that the tools used during the evaluation process are as important as the evaluation dialogue corpora. We have developed a methodology for evaluation of spoken dialogue systems that can be summarized in the following steps:

1. Definition of scenarios and questionnaires
2. Field trials and user satisfaction evaluation through questionnaires
3. Transcription and annotation of evaluation dialogues
4. Extraction of evaluation metrics
5. Analysis of results

During this process of evaluation we have used several tools, not only for the process of transcription and annotation but for the automatic extraction of metrics; for example we have used tools like NB (NB, 1998), MATE (MATE, 1998), HResults from the HTK Toolkit (ENTROPIC, 2000), MATLAB (MATLAB, 2002), LT-XML (HCRC, 2000) and our own developed tool ULAT (Utterance Level Annotation Tool)(Húder, 1998). Some of them were used specifically for annotation of dialogues and some others for post-processing of dialogues, that is, for extraction of metrics or statistics and calculation of values like kappa, recognition percentage etc. As we have mentioned before we have used several tools which where applied in different steps during our evaluation methodology, although we consider that it could be important to have available only one tool with which to carry out the very common evaluation procedures. In this paper we present an ongoing work of developing of such a tool, a tool that not only allow us transcribe and/or annotate dialogues of evaluation, but it also enables us to extract and calculate automatically several evaluation metrics from the corpus annotated in XML language (XML, 1997). A preliminary version of this tool was used during the evaluation of two prototypes of dialogue systems at Telefónica I+D (Charfuelán et al., 2000a; Charfuelán et al., 2000b), which give us invaluable experience to make it corrections and enhance of its capabilities. In fact, in this paper we present a new version of this tool which we have called ULAT-STAT. ULAT-STAT because now it integrates a new module for extracting and calculating metrics and statistics of evaluation, nevertheless preserving the same general characteristics of the previous version:

- XML based, so it is independent of a particular annotation scheme.
- Implemented in free and open source software (C, Tcl/Tk), which makes it possible to extent and freely distribute it.
- Platform independent (Unix or Windows).

The goal of this paper is to explain our methodology of spoken dialogue systems evaluation as well as to show how the ULAT-STAT tool can facilitate most of the procedures usually applied in evaluation of SLDS (EAGLES, 1996; Walker et al., 2000; Brey et al., 2000; Minker, 1998), in particular in our methodology of evaluation. In Section 2 the main steps of our methodology of SLDS evaluation are briefly described. In Section 3 the main characteristics of the new ULAT-STAT are summarised. In Sections 4 and 5 the procedures of transcribing, annotating and extracting metrics and statistics are described. In Section 6 some comments about the usefulness of evaluation results are made and in Section 7 conclusions are presented.

2. Methodology of evaluation of SLDS

The proposed methodology includes not only a list of steps to follow but the formalising of procedures for annotation, representation and processing of dialogues of evaluation.

- Formalising of evaluation dialogues annotation, because a multilevel scheme of annotation has been defined in XML language.

[1]The authors were partially supported by the Natural Interactive Systems Laboratory (NISLab), University of Southern Denmark, Forskerparken 10 DK-5230 Odense M, Denmark

- formalising of representation, because a set of annotated evaluation dialogues and the corresponding audio recordings have been organised in a well structured database of dialogues of evaluation.

- Formalising of processing, because automatic procedures for extracting information (metrics, statistics) from the XML database of evaluation dialogues have been defined.

The steps to follow in the proposed methodology of SLDS evaluation are:

1. Definition of scenarios and questionnaires, the definition of the scenarios depends on the system tasks under evaluation and the questionnaires, that the user must complete after the field trial, are mainly related to the system performance.

2. Field trials and afterwards user satisfaction evaluation through questionnaires, in our last evaluation (Bel et al., 2002) this step consisted of enabling a telephone number which the users called to and a web page in which the users answered the survey questions.

3. Transcription and annotation of field trial dialogues, the log files and the corresponding audio files are used as starting point. The user interventions for each turn are transcribed, and two levels of annotation in XML language are created: utterance level in which users and system turns are transcribed and/or annotated; and dialogue level in which task or dialogue segments are annotated. The result of this step is an evaluation dialogue database. This step is described in a little more detail in Section 4.

4. Extraction of evaluation metrics, for example: number of tasks completed per user or average for all the tasks of a particular type contained in the database, percentage of turns of user or system per task, percentage of word recognition per task, user satisfaction per task. Almost all these metrics and others defined in the tool can also be calculated on average for all the tasks of a particular type contained in the database or for a particular set of users. For example in our last evaluation (Bel et al., 2002) the database of evaluation dialogues was divided in two sets, Spanish speakers, Catalan Speakers. This step is described in a little more detail in Section 5.

5. Analysis of results, the Paradise (Walker et al., 1998) framework was used as a preliminary method of analysis, specially to study the effect of each metric in the global performance of the system.

3. The ULAT-STAT XML tool

This tool has been already described in (Charfuelán et al., 2000a), for explanation purpose its main characteristics are summarised here, describing when necessary the new features include in the new version. The ULAT-STAT tool has been developed using Tcl/Tk programming language including the freely distributed package SNACK from KTH (KTH, 1997), which provides an easy way to access audio files. The main characteristics of the ULAT-STAT tool are:

- Manual transcription of user turns having a controlled access to the audio file.

- Automatic extraction of information related to system turns, recognizer and parser outputs, and subjective information of the user from log files and external information files. In the new version the inclusion of subjective information coming from the survey questionnaires is made almost automatically, this was possible because the web pages in which the users complete the questionnaires have already generated information in XML format.

- Inclusion of information (attributes at different levels) from a human evaluator or annotator, for example, whether or not the user's concept (dialogue act) is lost after the speech recognition and parsing process. This feature has also been improved in the last version because the dialogue, task or turn attributes that are going to be evaluated in a particular dialogue system can be defined and configured using the graphic interface.

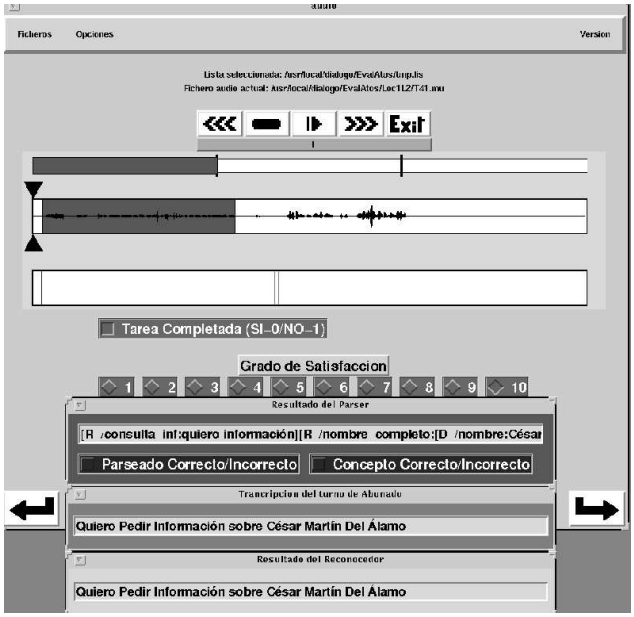

Figure 1: ULAT-STAT Annotation GUI

4. Transcribing and annotating with the ULAT-STAT tool

The inputs to the ULAT-STAT tool are: a recorded audio file of the dialogue, a log file provided by the dialogue system and external information (questionnaires). The graphic interface of the ULAT-STAT tool presents the information contained in the log file and audio file for each turn. During the annotation of a user turn, there appears three windows in the graphic interface, (Figure 1), they contain the recognized text, the parsed text and the transcription text. This last window appears filled with the recognised text, which

sometimes is the same or very close to what the person actually said (i.e. the correct transcription); in this case the annotator's tasks only consist of listening to the recording to verify its correctness. In that way the orthographic transcription of utterances becomes fast and easy.

4.1. Annotation Framework

This is another feature that has been changed in the last version of ULAT-STAT. The main difference between the previous annotation framework and the current one is that only one XML file is generated for each user involved in the field trial, basically the dialogue transcription (objective information) and the survey questionnaires (subjective information) are gathered in the same file in a well formed XML structure. An example of how the new XML structure looks like is showed in Figure 2

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE transcription SYSTEM "utterance.dtd">

<evaluation id="ATOS"
<transcription id="Thursday1A">
<task id="T11" completed="Yes" satisfaction="8">
  <phrase id="phr_1" who="system">
   <system id="sys_1" helps="No" >
    <wavsys id="wav_1" file="/Thursday1A.mu"
      start="2001" end="51001">
      Welcome, I am Atos an automatic
      telephone operating system,
      £What function do you want to do?.
    </wavsys>
   </system>
  </phrase>
  <phrase id="phr_2" who="user">
   <user id="user_1" corr="Yes" interrupt="No" shut="No"
     lost="No", helpu="No" fail_speech_detector="No">
    <wavusr id="wav_2" fich="/Thursday1A.mu"
       start="51001" end="57001">
      <trans id="trans_1">
       I want information about
       Cesar Martin Del Alamo
      </trans>
      <rec id="rec_1">
       I want information about
       Cesar Martin Del Alamo
      </rec>
      <par id="par_1" corr="Yes">
       [R_consult_inf: want information]
       [R_complete_name: [D_name: Cesar]
       [D_surname: Martin]
       [D_surname: DelAlamo]]
      </par>
    </wave>
   </user>
  </phrase>
   ...

 </task>
```

ULAT-STAT: Metrics per task, task 1 Catalan_set

User_No	Sat	Sat.prom	t_usu	t_usu_lost	C_corr	C_incorr	P_corr	t_sys	t_total	PWR	Duracion (seg)
1:	2	3.60	7	0	5	2	7	7	14	30.77	134.17
2:	4	2.10	6	0	2	4	6	5	11	16.67	80.34
3:	3	3.40	18	0	11	7	16	16	34	57.89	166.58
4:	3	3.00	18	1	8	10	18	18	37	40.43	246.22
5:	4	2.64	13	0	6	7	13	14	27	32.35	238.77
6:	3	3.30	24	0	12	12	23	18	42	61.22	168.90
7:	3	3.27	10	0	7	3	10	10	20	38.10	183.22
8:	2	3.60	14	1	8	6	12	10	25	26.09	235.23
9:	3	3.00	12	0	6	6	12	9	21	35.00	129.95
10:	3	3.60	6	0	5	1	6	6	12	78.95	123.80
Average	3.00	3.15	12.80	0.20	54.69	45.31	96.09	11.30	24.30	41.75	170.72

Figure 3: ULAT-STAT Metrics per task

- Sat: User satisfaction, which is the value of the question with which the user judges the global performance of the system.
- Sat_prom: User satisfaction average, which is calculated averaging the values of all the other questions with which the user judges different features of the system, like system friendless, system speech output (synthesis), system help etc.
- t_usu: Average of user turns per task.
- t_usu_lost: Average of user turns per task, in which the user is lost during the dialogue.
- C_corr: Average of user concepts that the system understood correctly, per task.
- C_incorr: Average of user concepts that the system did not understand correctly, per task.
- P_corr: Average of user turns per task, in which parsing was successful.
- t_sys: Average of system turns per task.
- t_total: Average of users and system turns per task.
- PWR: percentage of word recognition per task, this metric is calculated comparing what the user actually said and what the recogniser produced. The comparison is based on a dynamic programming sentences alignment procedure.
- Duration: Elapsed task time.

Using The ULAT-STAT tool the metrics can also be presented in graphs as curves, (Figure 4). Here the three metrics-curves are calculated per user and for same task, task 1. The metrics presented are: "PWR" percentage of word recognition, "shut" percentage of times the recognizer was triggered by noise and "fail_speech_detector" the percentage of times that the speech detector fails. The information presented in Figure 4 can also be presented as a graph of bars as we can see in Figure 5

Another kind of information that can be extracted from the database, not necessarily metrics, is the dialogue itself, because sometimes is very useful to analyse what happened during the interaction in a particular dialogue. In Figure 6 one particular dialogue of evaluation is shown, the information is presented turn by turn. The additional labels presented in curly brackets, in front of each turn, are annotated attributes, these attributes can be defined beforehand, so attributes can be added or removed from this screen.

6. Analysis and deployment of evaluation results

The methodology for evaluation of SLDSs and the XML-based tool presented in this paper have been tested during the evaluation of the E-MATTER prototype (Bel et al., 2002) E-MATTER is a multilingual Spoken Dialogue System (SDLS) designed to provide e-mail access over the telephone by combining different technologies such as continuous speech recognition, text-to-speech conversion, semantic parsing, dialogue management, language identification and text verification. Therefore the evaluation of this system faced us with the challenge of managing different data and metrics for different highly specialized Natural Language Processing (NLP) modules. In fact a two level evaluation methodology was defined for both isolated modules and the global dialogue level. And for both levels several metrics were obtained from the same data extracted from XML-files. These files were generated by applying our methodology, and XML-based annotation tool, over a set of dialogues collected under two scenarios designed for testing the usability of the whole system. As an illustration of the possibilities of our evaluation environment we present some global results obtained under the E-MATTER evaluation. The detailed analysis of this evaluation can be found in (Bel et al., 2002)

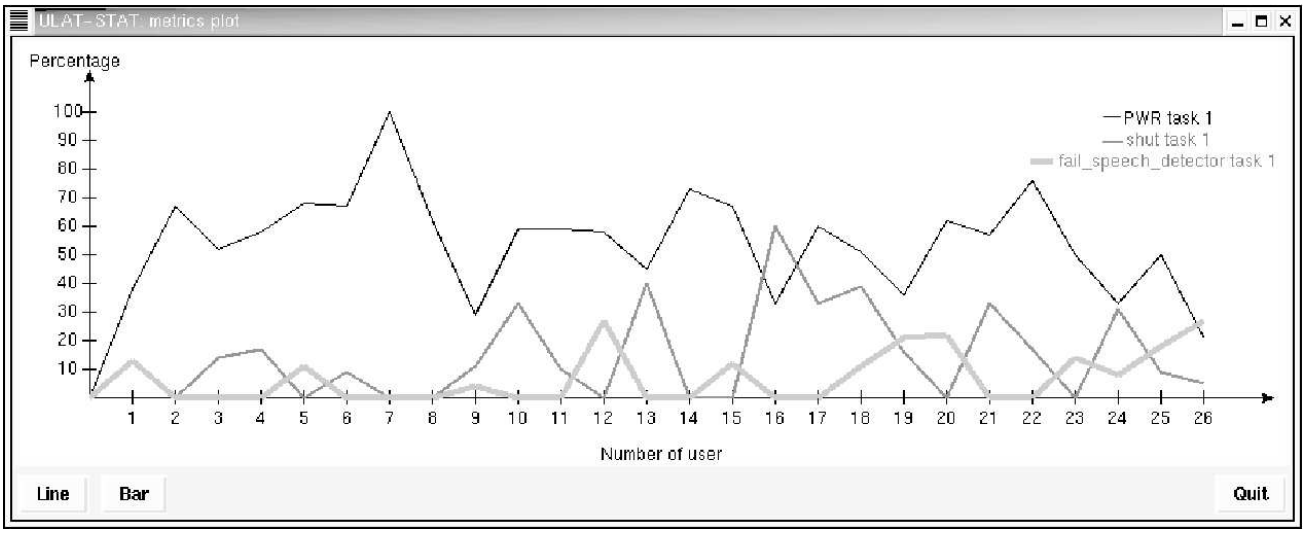

Figure 4: ULAT-STAT Metrics plot (line)

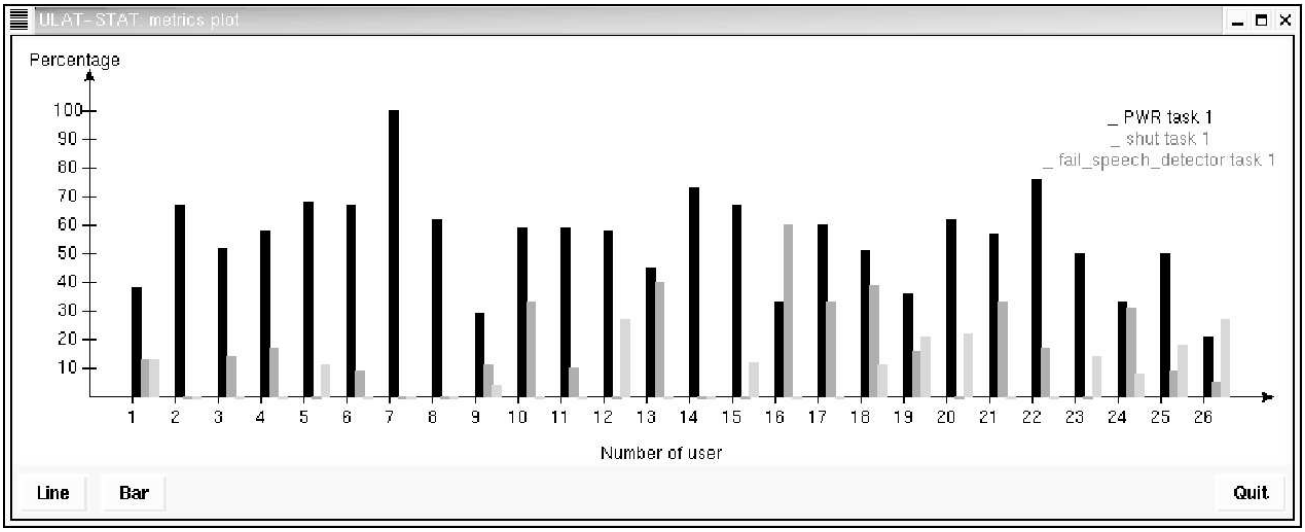

Figure 5: ULAT-STAT Metrics plot (bar)

Two simulated but realistic tasks or scenarios were defined: In the first one the user had to look for an e-mail from his/her office and then reply it using a predefined standard text message, while in the second scenario, the user was instructed to look for an e-mail where his/her friends informed about the date and place of a party, and then to reply with a voiced-recorded message. A population of 42 subjects was selected for the field trial, all test persons were novice users of E-MATTER, and several subsets of users were defined to evaluate some specific features of E-MATTER (for example, actually the system can be accessed both in Castilian Spanish and in Catalan, therefore two different subsets of users were established to compare their evaluation results). More detailed results and comparative analysis can be found in (Bel et al., 2002). In his work, trying only to give an overview of the possibilities of our analysis and evaluation tools, we present some global results. Firstly, we have to point out that a fully transcribed and annotated dialogue database composed of 84 tasks (2 tasks per users, and approximately 4 hours of recordings) was generated in 9 labeling sessions of 2 hours each, by three different persons who annotate the files using the ULAT-STAT XML annotation tool.

Only small discrepancies were found on the criteria followed for annotating some subjective metrics, specially those turns where a user was supposed to be lost in the dialogue. Then automatic extraction of evaluation metrics was done by processing the XML annotated database. As we mentioned before two different levels of analysis (at module and global dialogue) followed the extraction of evaluation metrics. A module evaluation level was followed for the evaluation of specific modules such as the Speech Recognizer and the Semantic Parser. As an example, Table 1 presents several evaluation metrics obtained for 42 dialogues corresponding to task 1 (mail from office) and 42 to task 2 (mail from friends). No distinctions between Castilian Spanish and Catalan speakers was made for generating the data presented in table 1, although this detailed analysis

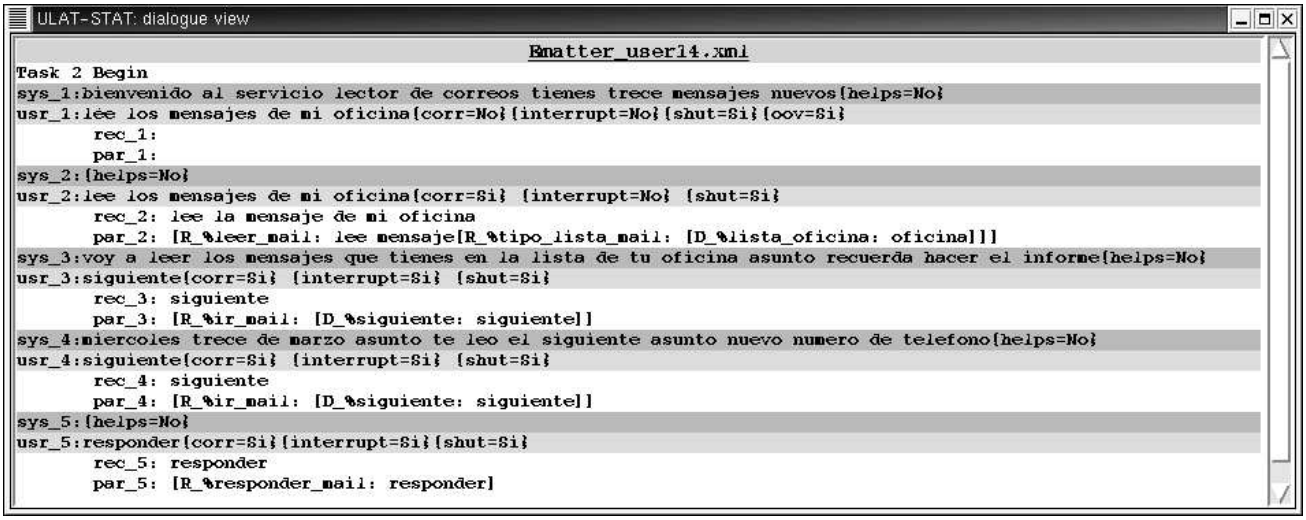

Figure 6: ULAT-STAT Dialogue box

ASR and Parser metrics	Task 1	Task 2
WER	50	42.13
PER	2.75	3.6
CA	66.8	68.9
Av. no. Turns:		
Barge-in	44.19	46.85
FalseRec	15.5	9.4
NoDetect	8.3	9.3
RecDisabled	4	5.5

Table 1: Speech recogniser and parser evaluation

Metrics	Task 1	Task 2
ET (seg)	173.1	159.7
UT	14.3	13.1
ST	11.1	9.8
Timeouts %	2.8	5.0
Helpu %	1.1	0.44
Helps %	3.4	1.4
UserLost	3.9	4.5
Comp	34	34
US	39.12	39.12

Table 2: Performance measures means

WER:	Word Error Rate
PER:	Parser Error Rate
CA:	Concept Accuracy.
Metrics as average number of turns:	
Barge-in	recognition under barge-in
FalseRec	false recognition
NoDetect:	no speech detection
RecDisabled:	attempts of the user to interrupt the system when the recognizer is disabled

can be found in (Bel et al., 2002)

Considering now the evaluation at the global dialogue level, we configured our ULAT-STAT tool to provide those metrics usually needed to follow the PARADISE evaluation framework (Walker et al., 1998). PARADISE tries to include and combine most of the proposed Spoken Dialogue evaluation procedures both from an efficiency and performance point of view. Several dialogue metrics for task success and both objective and subjective dialogue costs are selected to be combined performing a step wise multivariate linear regressions with user satisfaction (SAT) as the dependent variable. SAT is derived through a set of questions on different aspects of the users' perceptions of their interaction with the System. Therefore ULAT-STAT tool generates most of these metrics: Task success metric, the user perception of having accomplish the planned task, referred to as perceived task success (Comp), is also obtained from users' surveys.

Dialogue Efficiency: total elapsed time in seconds (ET) and number of system (ST) and user turns (UT). Dialogue Quality measures: time out prompts (Timeouts), number of user helps (Helpu), number of system helps (Helps), number of turns in which the user is lost during the dialogue (UserLost), user barge-in (BargeIn), concept accuracy (CA), false recognition (FalseRec), no speech detection (NoDetect) and attempts of the user to interrupt the system when the recognizer is disabled (RecDisabled).

In order to have the possibility of comparing different SLDSs, instead of raw counts it is usual to normalized (%) the quality metrics by dividing the raw counts by the number of utterances in the dialogue. As an illustration, Table 2 summarizes these dialogue metrics for task 1 and task 2 in the E-MATTER evaluation. Results for CA, BargeIn, FalseRec, NoDetect and RecDisabled, were given in Table I.

In Table 2 we have also included the subjective metrics: user perception of task success (Comp) and user satisfaction (SAT). Finally, and following the PARDISE framework, all the evaluation metrics can be used to train several models for different sets of dialogues corresponding to different user populations, tasks or SLDS behavior. As

Scenario	Factors	R^2
Task 1	NoDetect%, ET, FalseRec%	0.41
Task 2	FalseRec%, Timeout%, WER, CA	0.32

Table 3: Significant prediction factors

an example we performed stepwise multivariate linear regressions with user satisfaction as the dependent variable and the independent variables shown in Tables 1 and 2. An overall summary of our results for task 1 and task 2 is presented in Table 3, where we show which factors were found to be a significant predictors of user satisfaction, ordered by the degree of contribution. Table 3 also presents the variance in R^2, which gives an idea of the contribution of the combined factors to the variance of US, and is a descriptive measure of how strong is the linear association between metrics and user satisfaction.

7. Acknowledgments

This work has been partially founded by the EC project IST-1999-21042 and the I+D+I Spanish project TIC-1669-C04-4.

8. Conclusions

We have presented our latest improvements and applications of a XML-based tool aimed for formalising procedures for annotating, representating and processing of evaluation dialogues. Together with the presentation of a general annotation methodology, we have extended the use of the ULAT-STAT XML tool from transcribing and annotating to the extraction of different quality metrics and statistics usually found in the context of different SLDLs evaluation approaches. As a test bed for the XML tool and the evaluation methodology, we have also presented some evaluation results from its use on a two level evaluation, defined for both isolated modules and the global dialogue level, for the SLDS E-MATER. As a final remark, it is important to emphasize that the availability of a XML annotated dialogue database and a systematic evaluation methodology should be of great value during the development cycle of a SLDS. Of course the dialogue evaluation results can provide important information to detect and correct deficiencies both in some particular modules, and in the design of the discourse structure and control strategies. But only properly annotated dialogues, as those provided by the ULAT-STAT XML tool, will allow us to have an easy access to particular dialogues, or turns, where these specific problems occur, and then use case-base analysis to try to solve them. For example, the identification of the speech acts related to those turns where the Speech Recognizer or the Semantic Parser presents their lower performance rates can direct us to analyze the language models or parser grammars used to model or represent these dialogue acts.

9. References

Nuria Bel, Javier Caminero, Luis Hernández Gómez, Montserrat Marimón, José F. Morlesín, Josep M. Otero, José Relaño Gil, M. Carmen Rodríguez, Pedro M. Ruz, and Daniel Tapias. 2002. Design and evaluation of slds for e-mail access through the telephone. In *Proceedings of Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

Thomas Brey, Gerhard Hanrieder, Paul Heisterkamp, Ludwig Hitzenberger, and Peter Regel-Brietzmann. 2000. Issues in the evaluation of spoken dialogue systems - experience from the access project. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens. Greece.

Marcela Charfuelán, José Relaño Gil, M. Carmen Rodríguez, Daniel Tapias Merino, and Luis Hernández Gómez. 2000a. Dialogue annotation for language systems evaluation. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens. Greece.

Marcela Charfuelán, Cristina Esteban López, José Relaño Gil, M. Carmen Rodríguez, and Luis Hernández Gómez. 2000b. A general evaluation framework to asses spoken language dialogue systems: Experience with call center agent systems. In *First workshop on Robust Methods in Analysis of Natural Language Data, ROMAND 2000*, Laussane, Swiss.

EAGLES. 1996. Expert advisory group on language engineering standards. http://coral.lili.uni-bielefeld.de/ gibbon/EAGLES/.

ENTROPIC. 2000. HTK hidden markov models toolkit. http://htk.eng.cam.ac.uk.

HCRC. 2000. LT-XML version 1.2 language technology group human communication research centre edinburgh university. http://www.ltg.ed.ac.uk/software/xml/index.html.

Gonzalo López Barajas Húder. 1998. Sistema integral de etiquetado de diáálogos. In *Proyecto Fin de Carrera*, Univ. Politecnica de Madrid.

KTH. 1997. Snack and tcl/tk scripting language. http://www.speech.kth.se/snack/.

MATE. 1998. Multi-level annotation tools engineering project overview. http://mate.nis.sdu.dk/.

MATLAB. 2002. MATLAB the mathworks. http://www.mathworks.com/products/matlab/.

Wolfgang Minker. 1998. Evaluation methodologies for interactive speech systems. In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada Spain.

NB. 1998. NB discourse annotation tool. http://www.theredesign.com/Technology/Dialogue/.

M.A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*, 12:317–347.

Marilyn Walker, Lynette Hirschman, and John Aberdeen. 2000. Evaluation for darpa communicator spoken dialogue systems. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens. Greece.

XML. 1997. Extensible markup language (xml). http://www.w3.org/XML/.