# **Dialogue Annotation for Language Systems Evaluation**

## Marcela Charfuelán, José Relaño Gil, M. Carmen Rogríguez Gancedo, Daniel Tapias Merino<sup>1</sup>, Luis Hernández Gómez

Dep. SSR ETSIT-UPM Ciudad Universitaria Madrid (Spain) {marcela,jrelanio,mcarmen,luis}@gaps.ssr.upm.es

<sup>1</sup> Speech Tecnology Group, Telefónica Investigación y Desarrollo , S.A. C. Emilio Vargas, 6 28043 Madrid (Spain) daniel@craso.tid.es

#### Abstract

The evaluation of Natural Language Processing (NLP) systems is still an open problem demanding further research progress from the research community to establish general evaluation frameworks. In this paper we present an experimental multilevel annotation process to be followed during the testing phase of Spoken Language Dialogue Systems (SLDSs). Based on this process we address some issues related to an annotation scheme of evaluation dialogue corpora and particular annotation tools and processes.

### 1. Introduction

This paper addresses one of the more recent NLP research areas of managing dialogue annotation schemes and data formats during the testing phase of Spoken Language Dialogue Systems. The Spoken Language community has made significant progress towards this goal (Walker et al., 1998; Price et al., 1992; Minker, 1998) and most of the proposals for spoken dialogue evaluation are based on the use of information from properly designed evaluation dialogue corpora. Generally these corpora are extracted from log files as the evaluated system is working, and no specific nor standardized annotation procedures are used to represent the relevant information (recent works are reported on (DARPA, 1999; Isard et al., 1998; Dybkjoer et al., 1998)).

Therefore, from our point of view, two major problems arises: i) For all those important scientific efforts to design standard evaluation procedures, it is very difficult, if not impossible, to test a same SLDS with different evaluation approaches. And ii) a huge amount of information that could be used for future adaptations of the system or that could be useful for other research groups is lost.Being aware of the importance of having annotated dialogue corpora, specially for the future development of standardized SLDS evaluation procedures and tools, we summarize our experience on this context in three main aspects:

*1. An annotation scheme* that, starting from log files, the corresponding audio files and the questionnaires answered for the people involved in the evaluation, provides multiple annotation levels for:

- orthographic transcriptions for user and system turns including time marks.
- recognizer's and parser's outputs for each turn.
- task delimitation (or dialogue segments when a dialogue includes more than one task or function).
- and some extra attributes added to each turn and task for correctness, completion and user satisfaction.

We use XML as basic coding language and propose some DTD files for our particular task.

2. An annotation tool for the inclusion of manual labels (time marks and labels for the different turns and dialogue segments) and additional information in the multilevel XML annotation scheme. The tool let us:

- transcript the user's turns having a controlled access to the audio file.
- include the objective evaluation of the human annotator about the parsers analysis of the recognizers output.
- also we can annotate the correctness of the concept (speech act) finally obtained in each turn and the completion and user satisfaction for each task.

The other information, system turns and recognizer's and parser's outputs, is extracted automatically by the tool from the corresponding log file.

*3. Automatic extraction of dialogue metrics* from the annotated corpora. This is an important point to be considered when designing the sequence of annotation steps. Walker et al. classify most of the commonly proposed dialogue metrics as objective or subjective.

- Objective metrics can be calculated without recourse to human judgment, and in many cases, can be logged by the spoken dialogue system so that they can be calculated automatically (i.e. number of turns or utterances or mean system response time).
- Subjective metrics require subjects using the system and/or human evaluators to categorize the dialogue or utterances within the dialogue along various qualitative dimensions (i.e. user satisfaction or percentage of contextually appropriate system utterances).

Therefore, in order to have a general annotation scheme suitable for most of the proposals for dialogue evaluation, the corresponding annotation process will be reliable for gathering objective and subjective information at utterance and dialogue segment levels. We now explain our annotation scheme and process in more detail and illustrate its

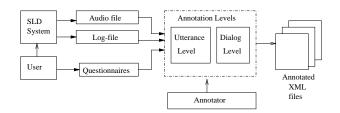


Figure 1: Block diagram of the global annotation process

application with our experience in the evaluation of an Automatic Telephone Information Service ATOS (Alvarez et al., 1996; Relaño Gil et al., 1999) developed by Telefónica I+D.

The goal of this paper is both to explain our annotation scheme and to illustrate its application with our experience in the evaluation of an Automatic Telephone Information Service SLDS. In Sections 2. and 3. we present our annotation process and scheme in some detail and illustrate with examples the multiple annotation levels. Section 4. describes the annotation tools with which we annotated dialogs. An example of practical results with the proposed scheme and annotation tools is presented in Section 5., in which we evaluate the ATOS system following the guidelines presented in this work. Finally in section 6. conclusions are made and future work is suggested.

### 2. Annotation Process

In this section we present the two annotation steps we follow in the annotation process for the development and evaluation of SLDS's. As it is shown in Figure 1, starting from three major sources of information: i) acoustic speech signal; ii) system log file; and iii) external information from the subjective evaluation of the user, we follow a two-step annotation process at utterance an dialogue level.

The first annotation step we define is at utterance level. At this level we perform two complementary tasks: to process logged information (for example: system response, recognizer's and parser's outputs) and to include all manual and/or subjective information (such as the transcription of user utterances, or whether the recognizer's or parser's outputs correctly captured the task-related information in the utterance).

The second annotation step is more global and it is defined at the dialogue level. Firstly, in this level some information related to the dialogue structure can be included (for example segments of dialogue corresponding to the starting and ending points of a particular task of the SLDS, or error recovery segments). After this dialogue structure mark-up is completed a set of simple automatic procedures are applied to obtain all the dialogue metrics and statistics needed for the evaluation of the SLDS. The output of the complete process is stored in annotated XML files as it is depicted in Figure 1.

### **3.** Annotation Framework

The annotation process described in the previous section will generate several annotation files for each dialogue having a XML structure like the example shown below: (for a description of XML Markup Language see (XML, 1997)).

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE transcription SYSTEM "utterance.dtd">
<transcription id="Thursday1A">
<task id="T11" completed="Yes" satisfaction="8">
  <phrase id="phr_1" who="system">
   <system id="sys_1">
    <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">
    <wavusr id="wav_2" fich="/ThursdaylA.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>
</transcription>
```

Main information in this annotated dialogue file include:

- The <transcription> element that is used as a global dialogue identifier.
- The <task> element, tagged at dialogue level, used to identify particular tasks or dialogue segments where specific evaluation metrics have been added.
- The <phrase> element obtained as a result of the annotation process at utterance level. It encloses all the elements related to user and system turns transcription, including:
  - Tags to link text information (transcriptions) to audio files information through <wavsys> and <wavusr> elements for the system and user audio turns respectively.
  - <trans> <rec> and <par> elements, enclosed by <wavusr> to gather information related to the evaluation of the speech recognition and semantic parser modules of the SDLS. <trans> contains the actual transcription of each user turn. <rec> and <par> include the output from the speech recognizer and the semantic parser with a specific attribute to indicate the presence of recognition and/or parser analysis errors.

Depending on the evaluation strategy, each file can contain one or several dialogue segments according with the number of tasks considered in an evaluation session, in the examples shown we evaluate several tasks for each transcription file. The corresponding and proposed DTD file for annotated evaluation dialogue files is as follows:

Evaluation Dialogues DTD "utterance.dtd"					
	erance.utu -	>			
ELEMENT</td <td colspan="5">transcription (task phrase)*&gt;</td>	transcription (task phrase)*>				
	transcription				
	id ID #REQUIRED >				
ELEMENT</td <td colspan="5">task (phrase)*&gt;</td>	task (phrase)*>				
ATTLIST</td <td colspan="5"></td>					
	id	ID	#IMPLIED		
	completed		#IMPLIED		
	satisfactior	1 (1 2 3 4 5			
		6 7 8 9 10)	#IMPLIED		
>					
ELEMENT<br ATTLIST</td <td>phrase (syst</td> <td>cem user)&gt;</td> <td></td>	phrase (syst	cem user)>			
AIILISI</td <td>id</td> <td>ID</td> <td>#REQUIRED</td>	id	ID	#REQUIRED		
	who	(system user)			
>	w110	(bybeem   uber /	#REQUIRED		
	system (wavsys)>				
ATTLIST</td <td></td> <td>1 ,</td> <td></td>		1 ,			
	id	ID	#REQUIRED		
>					
	wavsys (#PCI	DATA)>			
ATTLIST</td <td>wavsys</td> <td></td> <td></td>	wavsys				
	id	ID	#REQUIRED		
	file	CDATA	#REQUIRED		
	start	CDATA	#REQUIRED		
>	end	CDATA	#REQUIRED		
	NT ugor (wawugr)>				
ATTLIST</td <td colspan="5">user (wavusr)&gt;</td>	user (wavusr)>				
	id	ID	#REQUIRED		
	corr	(Yes No)	#REQUIRED		
>					
	<pre>wavusr (trans,rec,par)&gt;</pre>				
ATTLIST</td <td colspan="5">wavusr</td>	wavusr				
	id	ID	#REQUIRED		
	file	CDATA	#REQUIRED		
	start	CDATA	#REQUIRED		
	end	CDATA	#REQUIRED		
> ELEMENT trans (#PCDATA)					
ATTLIST</td <td></td> <td>117/2</td> <td></td>		117/2			
	id	ID	#REQUIRED		
>			~		
ELEMENT</td <td colspan="5">NT rec (#PCDATA)&gt;</td>	NT rec (#PCDATA)>				
ATTLIST</td <td></td> <td></td> <td></td>					
	id	ID	#REQUIRED		
>					
	T par (#PCDATA)>				
ATTLIST</td <td>-</td> <td></td> <td></td>	-				
	id	ID (Ver Ne)	#REQUIRED		
>	corr	(Yes No)	#REQUIRED		
-					

In the next example (below) we show how we can gather several annotated dialogue files, like the previous example, in a complete data base of evaluation dialogue files. We propose a structure in which the element <system> whose attribute is the name of the data base, collect several <dialog> elements. The <dialog> elements have a reference attribute to each of the transcription files. In this example we have collect only two files through referring the attribute "id" of each <transcription> element; this is a so simple example but we can in the same way (referring tags) extract from all the data base other important elements to our evaluation or study.

<!-- "EvalAtos.xml" -->

```
<!DOCTYPE eval SYSTEM "EvalAtos.dtd"
<!ENTITY fich_jlA "ThursdaylA.mu.xml">
<!ENTITY fich_j2B "Thursday2B.mu.xml">
1>
<system id="EvalAtos">
<dialog id="dl" href="&fich_jlA;#id(ThursdaylA)"/>
<dialog id="d2" href="&fich_j2B;#id(Thursday2B)"/>
</system>
<!-- DTD file to gather several dialogue files -->
<!-- "EvalAtos.dtd" -->
<!ENTITY % hrefAttr 'href
                                 CDATA
                                          #REOUIRED
                     xml:link
                                CDATA
                                          #FIXED "simple"
                     show
                                 CDATA
                                          #FIXED "embed"
                     actuate
                                CDATA
                                          #FIXED "auto"'>
<!ELEMENT system (dialog)*>
<!ATTLIST system
          id ID #REQUIRED
<!ELEMENT dialog
                  ANY>
<!ATTLIST dialog
          id ID #REQUIRED
          %hrefAttr;
```

<?xml version='1.0' encoding='ISO-8859-1'?>

In the following example we show the XML file to annotate the subjective information of the evaluation collected through the survey questions. Each <question> element has an attribute of identification an another of description. Afterwards the <eval\_question> elements contain the answers for each question obtained from each user involved in the evaluation. Here we again use reference attributes to the corresponding transcription files.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE Global-EvalAtos SYSTEM "Global-EvalAtos.dtd" [
<!ENTITY fich_j1A "Thursday1A.mu.xml">
<!ENTITY fich_j2B "Thursday2B.mu.xml">
1>
<evaluation id="EvalAtos">
            <!-- Survey Questions -->
<question id="Q1"
description= "Level of comprehension of the sys-
tem prompts"/>
<question id="Q2"
description="Frequency in which you can't follow the di-
alogue"/>
<question id="Q3"
 description="Level of comprehen-
sion of user turns in the dialogue"/>
<question id="Q4"
 description="At which level the system be-
come slow in its response time"/>
<guestion id="05"
 description="Give us a score (1 to 10) of global evalu-
ation of the system"/>
      <!-- Values for each question for each dialogue
->
<eval_question id="EV_1" href="&fich_j1A;#id(Thursday1A)">
 <eval question_id="Q1" value="7"/>
 <eval question_id="Q2" value="1"/>
 <eval question_id="Q3" value="9"/>
 <eval question_id="Q4" value="7"/>
 <eval question_id="Q5" value="7"/>
</eval_question>
```

<eval\_question\_id="EV\_2" href="&fich\_j2B;#id(Thursday2B)"/> <eval\_question\_id="Q1" value="10"/>

```
. . .
```

<sup>&</sup>lt;!-- XML file to gather several dialogue files -->

The next scheme shows the DTD structure of the previous XML file.

```
<!-- Global evaluation DTD file
     "Global-EvalAtos.dtd"
<!ENTITY % hrefAttr 'href
                                 CDATA
                                          #REOUIRED
                     xml:link
                                 CDATA
                                          #FIXED "simple"
                      show
                                 CDATA
                                          #FIXED "embed"
                                 CDATA
                                          #FIXED "auto"'>
                     actuate
<!ELEMENT evaluation (question*, eval_question*)>
<!ATTLIST evaluation
          id
                        ID
                                #REOUIRED
<!ELEMENT question EMPTY>
<!ATTLIST question
                                #REQUIRED
          id
                        ТD
                       CDATA
                                #REQUIRED
          description
<!ELEMENT eval_question (eval*)>
<!ATTLIST eval_question
          id
                                #REOUIRED
                        ID
          %hrefAttr;
<!ELEMENT eval EMPTY>
<!ATTLIST eval
          question_id
                        IDREF
                                                 #REOUIRED
                        (1|2|3|4|5|6|7|8|9|10)
          value
                                                 #REOUIRED
```

### 4. Processing Methodology and Tools

In the previous Section we have described how information from different sources are integrated into a dialogue annotation framework. We have also presented the global annotation strategy we propose following two sequential annotation steps at what we have referred to as utterance level and dialogue level. Now we discuss the annotation methodology which combines manual procedures from a human annotator with automatic processing of previously annotated data. This methodological approach allows us to have interactive feedback between different annotation criteria and their particular dialogue metrics results. Thus for example we could see on-line the possible impact of the speech recognizer for different segments in a dialogue. In subsections 4.1. and 4.2. we will describe the annotation process and related tools for the first annotation step at utterance level and the second annotation step at the dialogue level respectively.

#### 4.1. Annotation process and tools at utterance level

One of the main characteristics of the first annotation stage at utterance level is the requirement of an easy access to the audio speech file. For this we have developed an annotation tool for the utterance level that we call ULAT (Utterance Level Annotation Tool). ULAT has been developed under Tcl/Tk programming language including the freely distributed package SNACK from KTH (KTH, 1997), that provides an easy way to design a proper audio interface.

We can summarize the main processing stages at utterance level as the following ones:

• Manual transcription of user's turns having a controlled access to the audio file.

- Automatic extraction of information related to system turns, recognizer's and parser's outputs, and subjective information of the user from log files and external information files.
- Inclusion of subjective information from a human evaluator or annotator, for example, whether or not the user's concept (dialogue act) is lost after the speech recognizer and parser analysis.



Figure 2: ULAT Annotation Graphic Interface

The inputs to the ULAT tool are: a recorded audio file of the dialogue, a log file provide by the system and an external information file (questionnaires). Figure 2 is a view of the graphic interface of the ULAT tool and shows all the information of a user turn. The annotator only has to mark a section of the speech wave, corresponding to a turn, listen to it and check if the information that the ULAT tool has presented for the turn is correct. The beginning and end samples are obtained by the tool and recorded automatically in the output file. Additional features of the ULAT tool are:

- During the annotation of a user turn, there are three windows in the graphic interface containing the recognized text, the parsed text and the transcription text. This last window appears with the recognized text, which sometimes is the same or very close to what the person really said (i.e. the correct transcription); in this case the annotator's tasks consist only of listen to the recording to verify its correctness. In that way the orthographic transcription of utterances become fast and easy.
- The window that contains the parser's output has two buttons: One to indicate if after the recognition process and parser analysis the concept or data was lost, and the other to indicate if the parser analysis was correct.

#### 4.2. Annotation process and tools at dialogue level

Because the files generated by the ULAT tool are in XML format they can be processed directly, previous elaboration of stylesheet files, by the MATE Workbench (MATE, 1998) or by a great variety of available XML tools at different SLDS research centers. We made a stylesheet to annotate tasks (dialogue segments) an add the attributes of correctness and completion of these segments. Optionally we also use the query processor of the MateWorkbench to generate metrics and statistics (this capability of MateWorkbench is still under revision).

In this way we reach one of our objectives: to develop a scheme to annotate evaluation dialogues suitable to automatic extraction of different parameters and metrics like those suggested in PARADISE framework.

#### 5. Annotation and Evaluation Example

We now present our experience with the dialogue annotation scheme and related annotation tools that we have previously described, for the evaluation of an Automatic Telephone Operator System (ATOS) (Alvarez et al., 1996). The ATOS system provides simple telephonic services like: automatic phone call, multi-conference, voice messaging, automatic telephonic directory etc. The methodology that we have followed for the evaluation process is:

- Test scenarios definition and preparation of questionnaires for the subjects involved in the evaluation.
- Testing the system following the scenarios and completing the questionnaires. Thus audio files, log files and external information files are generated at this test stage.
- Annotation of recorded dialogues at utterance level.
- Annotation at dialogue level: dialogue segment tagging, addition of recognition performance information, and dialogue metrics for evaluation according to the PARADISE framework.
- Analysis of evaluation results.

The evaluation of our ATOS SLDS was done by selecting a subset of twelve different telephonic functions or tasks. A different identification code was assigned to each task. In order to have a proper preliminary evaluation of the system, a population of 30 subjects was selected, all of them were novice users of the ATOS system. Thus every subject involved into the evaluation process was informed on the basic functionality of ATOS and on the evaluation procedure.

As an illustration of the possibilities of using the final annotated files for SLDSs evaluation, we present some results for ATOS using the information stored in the Dialogue Database EvalAtos of XML annotated files. This results were obtained applying dialogue metrics for each task in ATOS, but we could also obtain evaluation results for the global behaviour of the system.

Table 1 contains dialogue metrics for the global behaviour of ATOS taking into account all the different tasks used for evaluation purposes. Here we have calculated the

Task	US	$\kappa$	TT	LT
T11	5.9200	0.3000	8.6400	3.2400
T12	6.1400	0.2222	8.0700	3.0500
T13	7.4200	0.4615	7.4200	2.1200
T41	5.6800	0.2727	8.4300	3.5900
T42	6.2800	0.3000	8.0700	2.7900
T51	5.9300	0.2500	13.200	3.8500
T52	3.5300	-1.166	11.690	4.5500
T53	6.3300	0.4231	16.260	4.6500
T61	5.8700	0.3846	10.250	3.5000
T62	5.2500	0.2727	13.870	4.3700
T71	5.5000	0.2000	16.810	4.6600
T81	4.6200	0.3333	16.620	7.7200

Table 1: Performance Metrics for a PARADISE Case Study: US = User Satisfaction,  $\kappa$  = Kappa coefficient, TT = Turns number for each Task, LT = Percentage of Lost Turns (turns with no correct concept)

 $\kappa$  coefficient as in (Walker et al., 1998) and we have chosen as cost measures the average number of turns for each task (TT) and the percentage of lost turns (LT), in order to do an analysis of variance (ANOVA) and estimate a performance function. (The analysis of results of this particular system can be seen in (Relaño Gil et al., 1999)).

### 6. Conclusions

In this paper we have presented a general framework for dialogue annotation in the context of the evaluation of Spoken Language Dialogue Systems. Our emphasis has been on two major issues: The design of a simple annotation scheme suitable for the wide range of dialogue evaluation methodologies, and a general annotation process implemented through XML coding and the use of standard tools and programming languages.

The resulting annotation framework is quite open and it can be easily configured or adapted to different approaches for SLDS assessment.

As we can see there are file references among different files, which is important in this kind of data base of multiple levels of annotation.

Following the annotation process both automatic and manual procedures have been presented to include objective and subjective data both from the user who is testing the system and from the human annotator. To examine the viability of the proposed coding scheme and annotation tools, they are being tested while evaluating a real SLDS using different dialogue metrics under the PARADISE framework. Although high flexibility is presented in this particular case, it is obvious that more evaluations of different SLDS are needed.

#### 7. References

- J. Alvarez, J. Caminero Gil, C. Crespo Casas, and D. Tapias Merino. 1996. The natural language processing module for a voice asisted operator at telefónica i+d. In *ICSLP* '96, Philadelphia, USA.
- DARPA. 1999. Darpa communicator log standard version history. http://fofoca.mitre.org/logstandard/.

- L. Dybkjoer, N. O. Bernsen, R. Carlson, L. Chase, N. Dahlback, K. Failenschmid, U. Heid, P. Heisterkamp, A. Jonsson, H. Kamp, I. Karlsson, J. v. Kuppevelt, L. Lamel, P. Paraubek, and D. Williams. 1998. The disc approach to spoken language systems development and evaluation. In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada Spain.
- Amy Isard, David McKelvie, and Henry S. Thompson. 1998. Towards a minimal standard for dialogue transcripts: A new sgml architecture for the hcrc map task corpus. In *Proceeding of International Conference on Spoken Language Processing*, Australia.
- KTH. 1997. Snack and tcl/tk scripting language. http://www.speech.kth.se/snack/.
- MATE. 1998. Mate. project overview. http://mate.nis.sdu.dk/.
- Wolfgang Minker. 1998. Evaluation methodologies for interactive speech systems. In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada Spain.
- José Relaño Gil, Daniel Tapias, M. Carmen Rodríguez, Marcela Charfuelán, and Luis Hernández Gómez. 1999. Robust and flexible mixed-initiative dialogue for telephone services. In Ninth Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway. Proceedings of EACL '99.
- Patti Price, Lynette Hirschman, Elisabeth Shriberg, and Elizabeth Wade. 1992. Subject-based evaluation measures for interactive spoken language systems. In DARPA Proceedings of Speech and Natural Language Workshop.
- 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.
- XML. 1997. Extensible markup language (xml). http://http://www.w3.org/XML/.