Chapter 2: Evaluation of Spoken Dialogue Systems

2.1 Subjective Evaluation

In order to evaluate different aspects of the quality of a spoken-dialogue-system-based service, subjective experiments with human users have to be carried out. According to the ITU-T Recommendation (page 851), such experiments serve two main purposes:

1) “During the interaction, instrumentally measurable system variables are collected, and the utterances of the system and the user are logged. The log-files are submitted to an expert evaluation, the outcome of which is a set of variables describing specific aspects of the human-machine interaction on the utterance, dialogue and task level, from a system developer's point of view. “

2) “After the interaction, test subjects are given a questionnaire that aims at collecting information about the perceptive quality features which are relevant to form the overall quality impression of the human user. Such experiments can be performed with fully functional systems, or with systems which are still in the development phase and where parts of the system modules have to be simulated. Details on the experimental set-up, the questionnaires, and on usability evaluation methods are given in clauses 6 to 8.”

This means that the subjective measures, aimed at assessing the users' opinions on the system, are obtained through direct interview by questionnaire filling. Questions including issues such as ease of usage, naturalness, clarity, friendliness, robustness regarding misunderstandings and subjective length of the transaction.

Subjective experiments can either be carried out with fully working systems (like the INSPIRE system), or with the help of a human experimenter simulating missing parts of the system, or the system as a whole (a so-called "Wizard-of-Oz simulation", like the BoRIS System). In order to obtain valid and reliable results, the (simulated) system, the test users, and the experimental task have to fulfill several requirements, see clauses 6.1 to 6.3 from ITU-T Recommendation (page 851).

The quality evaluation values that will be used as dependent variables for all the prediction models are the following:

-Question B0 → Overall Quality (question 1.0 from the INSPIRE system questionnaire)

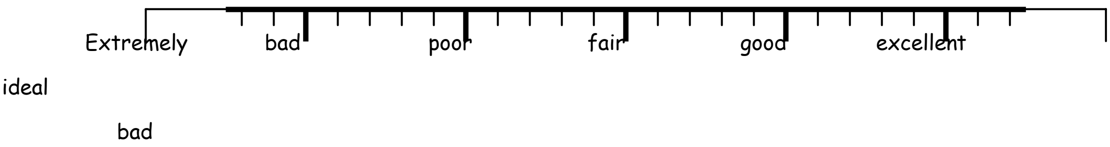
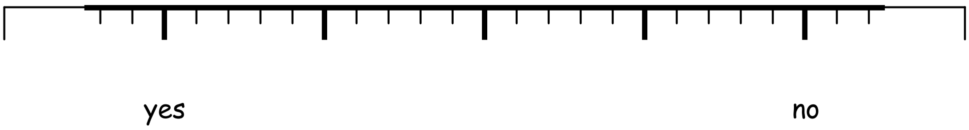


Figure 1: Bodden-Jekosch Scale with allocated concepts

-Question B23 → Overall user satisfaction

23. Overall, you are satisfied with the dialogue:



-"Mean Questions B" → A mean value of all questions from the Questionnaire B (chapter 8.1).

- "Use again" → For the INSPIRE system. Questions if the user would use the system again or not (question 7.2 from the INSPIRE questionnaire, see Chapter 8.2) :

7.7 I would use the System again in the future.

I strongly agree I agree Undecided I disagreeI strongly
disagree

□ □ □ □ □

-“Easy learning” → For the INSPIRE system. Questions if the user found that the way the system works was easily learned (question 7.7 from the INSPIRE questionnaire, see chapter 8.2):

7.2 The usage of the System was easy to learn.

I strongly agree I agree Undecided I disagreeI strongly
disagree

□ □ □ □ □

Each of the questions on the questionnaires that the user answers is a way of measuring subjective evaluation. For more details and the entire questionnaires from BoRIS and INSPIRE, see Chapter 8.

2.2
Parametric description of interaction

Interaction variables describe the characteristics of the system, the user and the interaction between them. Usually, it is not possible to separate the influences from the system and the user because the user’s actions are strongly influenced by the behavior of the system.

Quality perceived by the user can only be measured in a direct way by collecting user judgments in a laboratory or field test situation. Instrumentally or expert-derived variables may carry very useful information on the interaction between the user and the system.

Interaction variables may be calculated on a word, sentence, utterance or dialogue level.

The set of variables collected during the evaluation of a spoken dialogue system are related to:

- dialogue and communication ;
- meta-communication (i.e. communication about communication) ;
- cooperativity;
- task success;

- speech input;

These aspects have been identified as major contributing aspects to system usability, user satisfaction and acceptability; see Möller (2002, 2004).

2.2.1 Instrumentally-measured and expert-annotated variables

The interaction between the system and the user is based on a sequence of alternated turns taken from both parts, with questions, answers, propositions, confirmations or corrections. From this sequence of turns, interaction variables can be obtained. This extraction of the variables is either done on an instrumentally way (for instance the duration of the dialogue) or done by a human expert who does the transcriptions (for instance, appropriateness of system utterances, task success).

The following Table gives an overview of the instrumentally-measured variables collected for each dialogue between user and system:

Table 1: Interaction variables instrumentally measured during the experiments

Abbr.	***Name***	***Definition***
DD	dialogue duration	Duration of the dialogue (in seconds).
STD	system turn duration	Average duration of a system turn (in seconds), from the system starting speaking to the system stopping speaking.
UTD	user turn duration	Average duration of a user turn (in seconds), from the user starting speaking to the user stopping speaking.
SRD	system response delay	Average delay of a system response (in seconds), from the user stopping speaking to the system starting speaking.
URD	user response delay	Average delay of a user response (in seconds), from the system stopping speaking to the user starting speaking.
# turns	number of turns	Overall number of turns (count) uttered in a dialogue. A turn is an utterance, i.e. a stretch of speech spoken by one party in the dialogue.
# system turns	number of system turns	Overall number of system turns (count) uttered in a dialogue.
# user turns	number of user turns	Overall number of user turns (count) uttered in a dialogue.
WPST	words per system turn	Average number of words (count) per system turn.
#system words	number of system words	Overall number of system words (count) uttered in a dialogue.
WPUT	words per user turn	Average number of words (count) per user turn.
# user words	number of user words	Overall number of user words uttered in a dialogue (count).
#ASR rejections	number of ASR rejections	Overall number of ASR rejections (count) in a dialogue. An ASR rejection is defined as a system prompt indicating that the system was unable to "hear" or to "understand" the user.
#system questions	number of system questions	Overall number of system questions in a dialogue (count). A system question is defined as an explicit or implicit directive to the user to provide information to the system.
# system error messages	number of diagnostic system error messages	Overall number of diagnostic error messages from the system in a dialogue (count). An error message is defined as the indication from the system that the system is unable to perform a certain task.
# system help	number of diagnostic system help messages	Overall number of help messages generated by the system in a dialogue (count). A help message is a system utterance which informs the user about available options at a certain point in the dialogue.

The following diagram displays the instrumentally measured variables, their relation to one another and group them in terms of how they are measured:

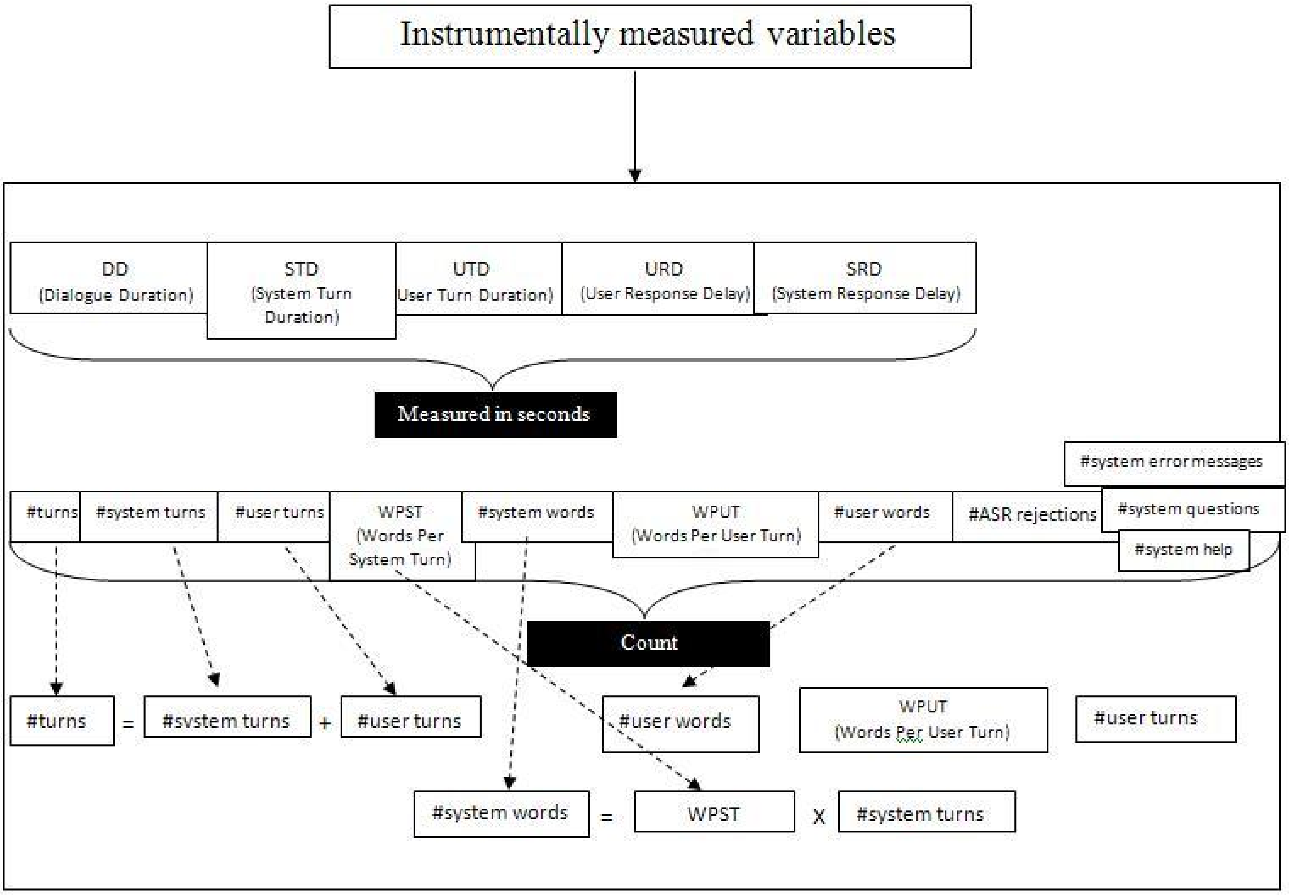

Instrumentally measured variables
DD (Dialogue Duration)
STD (System Turn Duration)
UTD User Turn Duration)
URD (User Response Delay)
SRD (System Response Delay)
Measured in seconds
#system error messages
#turns
#system turns
#user turns
WPST (Words Per System Turn)
#system words
WPUT (Words Per User Turn)
#user words
#ASR rejections
#system questions
#system help
Count
#turns = #svstem turns + #user turns
#user words
WPUT (Words Per User Turn)
#user turns
#system words = WPST X #system turns

Diagram 1 – Instrumentally measured variables

The following Table gives an overview of the variables collected by experts for each dialogue between user and system:

Table 2: Interaction variables collected by experts during the experiments.

Abbr.	***Name***	***Definition***
#user questions	number of user questions	Overall number of user questions uttered in a dialogue(count). A user question is labeled by the annotation expert
AN:CO, *AN:IN*, *AN:PA*, *AN:FA*	number of correct/ incorrect/ partially correct/ failed system answers	Number of questions from the user which are answered by the system, per dialogue (count): • correctly (*AN:CO*) • incorrectly (*AN:IC*) • partially correctly (*AN:PA*) • not at all (*AN:FA*)
DARPAs, *DARPAme*	DARPA score, DARPA modified error	Measures (in points) according to the DARPA speech understanding initiative, modified by Skowronek (2002) to account for partially correct answers: $DARPAs = \frac{AN:CO - AN:IC}{\#\ user\ questions}$ $DARPAme = \frac{AN:FA + 2 \cdot (AN:IC + AN:PA)}{\#\ user\ questions}$
#help requests	number of help requests from the user	Overall number of user help requests in a dialogue (count). A user help request is labeled by the annotation expert.
SCT, *%SCT*	number or percentage of system correction turns	Overall number (*SCT*) (count) or percentage (*%SCT*) of all system turns in a dialogue which are primarily concerned with rectifying a “trouble”, thus not contributing new propositional content and interrupting the dialogue flow. System correction turns are labeled by the annotation expert.
UCT, *%UCT*	number or percentage of user correction turns	Overall number (*UCT*) (count) or percentage (*%UCT*) of all user turns in a dialogue which are primarily concerned with rectifying a “trouble”,

Abbr.	***Name***	***Definition***
		thus not contributing new propositional content and interrupting the dialogue flow. User correction turns are labeled by the annotation expert.
# cancel	number of user cancel attempts	Overall number of user cancel attempts in a dialogue (count). A user cancel attempt is labeled by the annotation expert.
# barge-in	number of user barge-in attempts	Overall number of user barge-in attempts in a dialogue (count). A user barge-in attempt is labeled by the annotation expert.
CA:AP, *CA:IA*, *CA:TF*, *CA:IC*, *%CA:AP*, *%CA:IA*, *%CA:TF*, *%CA:IC*	contextual appropriateness	Overall number (count) or percentage of system utterances which are judged to be appropriate in their immediate dialogue context. Determined by labeling utterances according to whether they violate one or more of Grice's maxims for cooperativity: • *CA:AP*: Appropriate, not violating Grice's maxims, not unexpectedly conspicuous or marked in some way. • *CA:IA*: Inappropriate, violating one or more of Grice's maxims. • *CA:TF*: Total failure, no linguistic response. • *CA:IC*: Incomprehensible, content cannot be discerned by the annotation expert.
PA:CO, *PA:PA*, *PA:IC*	number of correctly/ partially correctly/ incorrectly parsed user utterances	Evaluation of the number of concepts (attribute-value pairs, AVPs) in an utterance which have been extracted by the system (count): • *PA:CO*: All concepts of a user utterance have been correctly understood by the system. • *PA:PA*: Not all but at least one concept of a user utterance has been correctly understood by the system.

Abbr.	*Name*	*Definition*
		• *PA:IC*: No concept of a user utterance has been correctly understood by the system. Expressed as the overall number of user utterances in a dialogue which have been parsed correctly/ partially correctly/ incorrectly.
IR	implicit recovery	Capacity of the system to recover from user utterances for which the speech recognition or understanding process partly failed. Determined by labeling the partially parsed utterances as to whether the system response was appropriate or not: $IR = \frac{\#\,utterances\ with\ appropriate\ system\ answer}{PA:PA}$
n_{AVP}, c_{AVP}, s_{AVP}, i_{AVP}, d_{AVP}, ot_{AVP}	number of identified semantic units	Overall number of semantic units (count) from all user utterances of a dialogue which have been • correctly understood (c_{AVP}) • substituted (s_{AVP}) • inserted (i_{AVP}) • deleted (d_{AVP}) • correctly not set (not_{AVP}) Determined from the overall number of concepts contained in all user utterances, n_{AVP}, by an expert annotation.
IC	information content	Percentage of correctly understood semantic units, per dialogue: $IC = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$
UA	understanding accuracy	Percentage of user utterances in which all semantic units (AVPs) have been correctly extracted: $UA = \frac{PA:CO}{\#\,user\ turns}$

Abbr.	***Name***	***Definition***
n, *c*, *s*, *d*, *i*	number of correctly identified/ substituted/ deleted/ inserted words	Overall number of words (count) from all user utterances of a dialogue which have been • correctly recognized (*c*) • substituted (*s*) • deleted (*d*) • inserted (*i*) Determined from the overall number of user words.
NEU	number of errors per utterance	Average number of recognition errors in an utterance (count). Being *s(k)*, *i(k)* and *d(k)* the number of substituted, inserted and deleted words in utterance *k*, then $NEU(k) = s(k) + i(k) + d(k)$ The average *NEU* can be calculated as follows: $NEU = \frac{\sum_{k=1}^{\#\ user\ turns} NEU(k)}{\#\ user\ turns} = \frac{WER \cdot \#\ user\ words}{\#\ user\ turns}$
WEU	word error per utterance	Related to *NEU*, but normalized to the number of words in utterance *k*, *w(k)*: $WEU(k) = \frac{NEU(k)}{w(k)}$ The average *WEU* can be calculated as follows: $WEU = \frac{\sum_{k=1}^{\#\ user\ turns} WEU(k)}{\#\ user\ turns}$
WER, *WA*	word error rate, word accuracy	Percentage of words which have been correctly recognized, based on the orthographic form of the hypothesized and the (transcribed) reference utterance. $WER = \frac{s + i + d}{n}$ $WA = 1 - \frac{s + i + d}{n} = 1 - WER$

Abbr.	***Name***	***Definition***
n_{iso}, c_{iso}, s_{iso}, d_{iso}	number of correctly identified/ substituted/ deleted/ inserted words (isolated word recognition)	Overall number of function words (keywords of the recognizer's vocabulary) (count) from all user utterances of a dialogue which have been • correctly recognized (c_{iso}) • substituted (s_{iso}) • deleted (d_{iso}) Determined in a similar way as *c*, *d* and *s*, but ignoring insertions due to the keyword-spotting approach (isolated word recognition metrics).
NEU_{iso}, WEU_{iso}	number of errors per utterance, word error per utterance (isolated word recognition)	Metrics similar to *NEU* and *WEU*, but determined on the function words only, ignoring insertions (isolated word recognition metrics).
WER_{iso}, WA_{iso}	word error rate, word accuracy (isolated word recognition)	Metrics similar to *WER* and *WA*, but determined on the function words only, ignoring insertions (isolated word recognition metrics).
TSw	Weighted task success	Weighted average task success of the dialogue, by assigning a value of • +1 to *S*, *SCs*, *SCu*, *SCsCu* and *SN* • 0 to *Fs* and *Fu* and calculating the arithmetic mean over all sub-tasks.

The following diagram displays the expert-annotated variables and group according to their purpose:

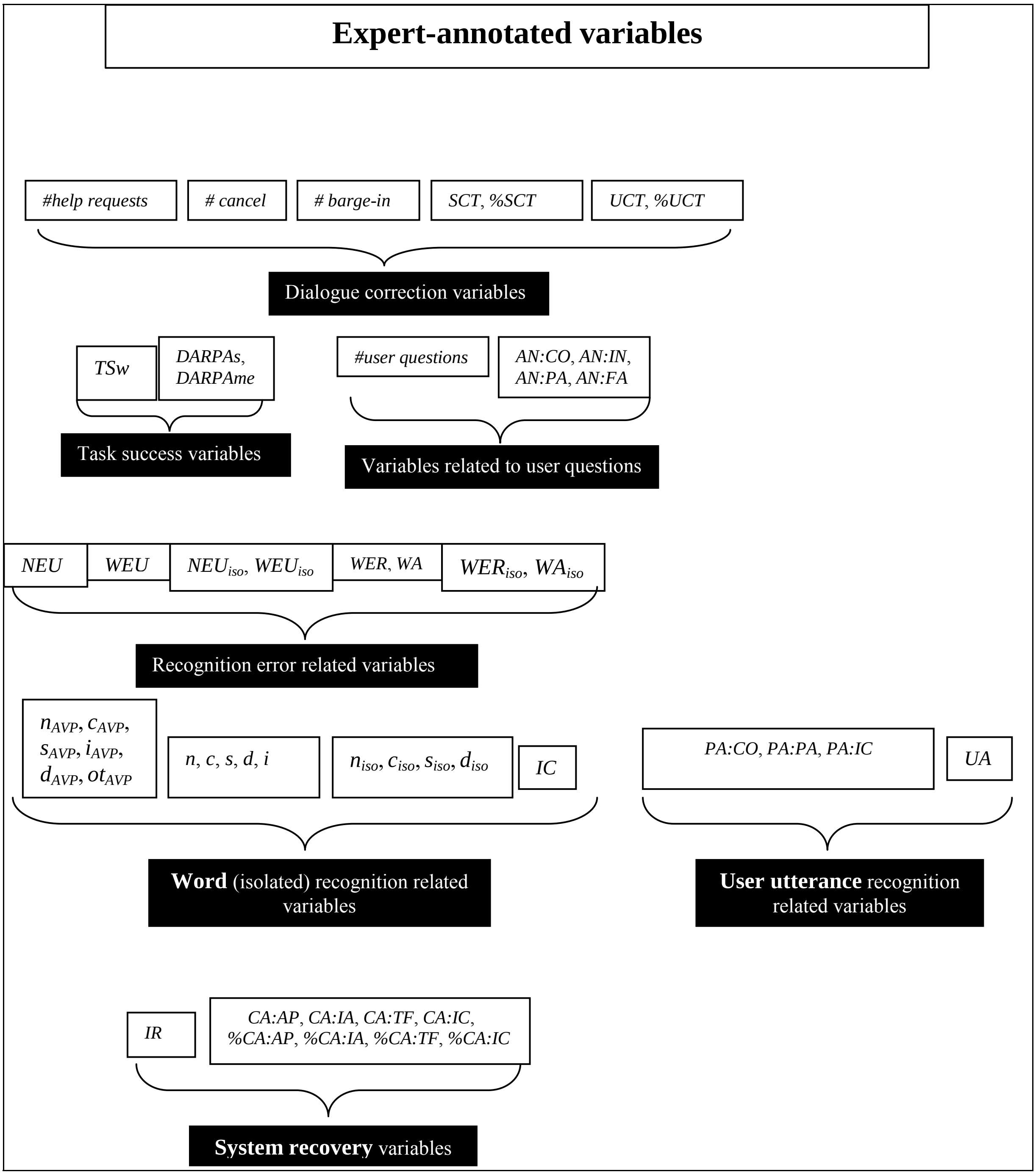

Expert-annotated variables
#help requests
# cancel
# barge-in
SCT, %SCT
UCT, %UCT
Dialogue correction variables
TSw
DARPAs, DARPAme
Task success variables
#user questions
AN:CO, AN:IN, AN:PA, AN:FA
Variables related to user questions
NEU
WEU
NEUiso, WEUiso
WER, WA
WERiso, WAiso
Recognition error related variables
nAVP, cAVP, sAVP, iAVP, dAVP, otAVP
n, c, s, d, i
niso, ciso, siso, diso
IC
Word (isolated) recognition related variables
PA:CO, PA:PA, PA:IC
UA
User utterance recognition related variables
IR
CA:AP, CA:IA, CA:TF, CA:IC, %CA:AP, %CA:IA, %CA:TF, %CA:IC
System recovery variables

Diagram 2 – Expert annotated variables

The variables DD, STD, STDlist, UTD, UTDlist, SRD, SRDlist, URD, URDlist, # turns, # system turns, # user turns, WPST, # system words, WPUT, # user words, # ASR rejections, # system questions, # system error messages and # system help, from Table 1 have been extracted directly from the log files generated by the dialogue manager. The other variables, from Table 2 require a transcription and annotation of the dialogue by a human expert.

It should be stressed that a new variable was developed in our study: weighted CA:IA. It has the same principle from CA:IA, but takes under consideration if the inappropriate system utterances were in succession or not. It is a squared sum of CA:IA´s basically (for example: 8 CA:IA´s total, but from these 8 times, 1 time 3 CA:IA´s in succession, 1 time 2 CA:IA´s in succession and 3 times 1 CA:IA alone → $3^2 + 2^2 + 3 = 16$ is the weighted CA:IA). The correlations of weighted CA:IA with the target variables are in Chapter 4.2 (selection of input and target variables).

Other variables have been considered for this work as well, but exclusively for the INSPIRE system. They are presented in details on the report “Error Coding for Free-Woz data”[1]. They classify the errors that happened during the dialogue, annotated by an expert. The classes of errors that were used during this work are classified in the following categories:

Table 3: Interaction variables (errors) collected by experts during the experiments.

Interaction Variables **(errors from the INSPIRE system)**	**Definition**
'no input'	Failing to issue a command during the response interval where the system expects it to be issued.
'capability'	Issuing a command for action that cannot be performed by the system because it does not possess that capability. It is possible to think of an extension to the system that would be able to perform the intended action
'state'	Issuing a command that is valid and progressive (in regard with the goal expressed in the task given to the user in the experiment) in one state of the dialogue, but *not* in the current one. The progressiveness criterion can be compromised in some cases. It should be marked as Unprogressive State Error then.
'vocabulary and grammar'	Issuing a command that would be valid if one word was changed to its synonym or the grammatical order of words was changed, without changing the vocabulary nor the meaning of the utterance.
'word error'	if a word was changed to a synonym or expression with same meaning. Example: Asking for "presenting" the message instead of "playing" it. This Error types can be divided into verb, noun and adjective
'modelling'	Issuing a command that would be valid if the system represented the word in a different way. If it is possible to imagine another kind of model/categorization of the word, this error would not emerge. (This should not be confused with the state error, wherein order errors are related to dialogue structure, not the word and vocabulary errors, wherein errors cannot be drawn back to modelling of the word.)

‘space’	spatial categorization error. The user refers to the space in a way that is not understandable by the system. Example: “Please turn on the lamp that is on the right from the table lamp” (This fails because the system does not have a model of the relative positions of the lamps.) (Note: This should not be confused with vocabulary error in the case the reference is made in just one word.) Example: “Please turn on the lamps on the right” – The system knows that there are two lamps on the right, but does not model their (common) relationship from the perspective of the user.
‘unprogressive state error’	the progressiveness criterion is loosened, i.e. the command has to be valid, but the corresponding AVPs have been acquired already. Example: *S: I understood ANY as kind of Program. From your choice there are several possibilities. Please say the number of the title of your choice from the list on the display.* *U: Program information.*
‘repetition’	The system repeats the same prompt (word to word or just the end part of it but meaning the same thing and being pragmatically the same prompt with same action alternatives (E.g., “Was kann ich fuer Sie tun” (“What can I do for you”) is an often repeated shorthand for more complex prompts.).