Annotations for Dynamic Diagnosis of the Dialog State[*]

Laurence Devillers, Sophie Rosset, Hélène Bonneau-Maynard, Lori Lamel
LIMSI-CNRS, BP 133
91403 Orsay cedex, France
{devil, rosset, hbm, lamel}@limsi.fr

Abstract

This paper describes recent work aimed at relating multi-level dialog annotations with meta-data annotations for a corpus of real human-human dialogs. This work is carried out in the context of the AMITIES project in which spoken dialog systems for call center services are being developed. A corpus of 100 agent-client dialogs have been annotated with three types of annotations. The first are utterance-level DAMSL-style dialogic labels. The second set of annotations applies to exchanges and takes into account of the dynamic aspect of dialog progress. Finally, 5 emotions types are annotated at the utterance level. Some of these multi-style annotations were used in a multiple linear regression analysis to predict dialog quality. The predictive factors are able to explain about 80% of the dialog accidents.

1. Introduction

Annotating dialog corpora has been and continues to be an active research area [1]. In the context of conversational dialog systems, annotations at multiple levels (lexical, syntactic, semantic, dialogic, ...) play an integral role in the development and evaluation of the constituent components and of the complete system, as well as in studying the communication process. We are also exploring how to annotate meta-data (such as emotion and topic) and to correlate them with dialog quality, progression, and success.

Most reported dialog success rates are based on human assessments of transactions, which are usually founded on both objective measures and subjective judgments. Dialog success can be measured according to different criteria, most notably, the task success rate and the level of customer satisfaction. Task success is usually simpler to assess and is given a binary rating (success/failure). Customer satisfaction is often measured indirectly via questionnaires (friendliness, ease of use) or by measures that can be correlated with efficiency (number of turns, repetitions, reparations, etc.) [4, 9, 10]. Client satisfaction is often also evident from the user's mood, which can to some extent be determined by the words employed or by prosodic features. Most of the research activities in automatic emotion detection concern prosodic features extractions [5, 7]. One of our long term objectives is to find a way to detect the user's mood during the interaction by using both lexical and prosodic cues, and in doing so be able to assess the quality of the ongoing dialog.

This papers aims to relate dialog annotations at different levels (lexical, pragmatic and dialogic) and with a meta-level annotation for emotion in a corpus of real human-human dialogs. This study is being carried out within the framework of the European and American project AMITIES (Automated Multilingual Interaction with Information and Services) [2]. We have annotated a spoken human-human dialog corpus of 100 dialogs with three types of annotations: dialogic labels, dialog progression axe labels, and emotion labels. The second set of labels take into account of the dynamic aspect of dialog [6]. The annotations were carried out separately to ensure the independence of the annotations. A predictive function of dialog quality is derived from the relative contributions of various factors extracted from dialogic, progression and emotion annotations. The most important predictors are determined via a principal components analysis. These measures are able to explain about 80% of the dialog problems.

In the next section, the agent-client dialog corpus is described. The following sections describe the annotation methodologies adopted. We then present the predictive functions and the results of the factor analysis, indicating the most relevant factors in predicting problematic dialogs.

2. Corpus

The dialogs are real agent-client recordings from a Web-based Stock Exchange Customer Service center. These recordings were made for purposes independent of this study, and have been made available for use in developing an automated call routing service within the context of the AMITIES project. The service center can be reached via an Internet connection or by directly calling an agent. While many of the calls are concerned with problems in using the Web to carry out transactions (general information, complicated requests, transactions, confirmations, connection failures), some of the callers simply seem to prefer interacting with a human agent. A corpus of 100 agent-client dialogs (4 different agents) in French has been orthographically transcribed and annotated. The dialogs cover a range of investment related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management (open, close, transfer, credit, debit) and Web questions and problems. Table 1 summarizes the characteristics of the corpus. There are 6241 speaker turns, 5229 sentences after excluding overlaps which are known to be frequent phenomena in sponta-

# agents	4	*# clients*	100
# turns/dialog	ave: 50	min: 5	max: 227
# words/turn	ave: 9	min: 1	max: 128
# words total	44.1k	*# distinct*	3k

Table 1: Characteristics of the corpus of 100 agent-client dialogs.

[*]This work was partially financed by the European Commission under the IST-2000-25033 AMITIES project.

Information:	
	Task, Out-of-Task, Communication
Forward Looking function:	
	Statement, Influence on Listener
Backward Looking function:	
	Understanding, Agreement, Answer, Response-to

Table 2: DAMSL dialog annotation levels and corresponding label types. The values for each label type are given in Tables 4 - 8.

Label	*#Turns*	*%Turns*
Statement	2087	33%
Influence on listener	1221	20%
Agreement	895	14%
Understanding	1385	22%
Answer	823	13%
Response-to	2812	45%

Table 3: Dialog annotation characteristics. Note that the levels are not exclusive, so that the total number can be larger than 100%

neous speech. The corpus contains a total of 44.1k words, of which 3k are distinct. The average dialog length is 50 utterances (min 5, max 227), the average sentence length is 9 words (min 1, max 128).

3. Dialogic annotations

The dialogic annotations were adapted from the DAMSL standard dialog acts. A detailed description of the basic DAMSL labels can be found in *"Coding Dialogs with the DAMSL Annotation Scheme"* [3]. The turn based annotations were entered using the XDML tool provided by the State University of New York, Albany a partner in the AMITIES project (see Figure 1). For this study the DAMSL annotations were limited to the dialogic levels: no semantic labels were annotated. The selected DAMSL dialog labels are applied at three main levels: **Information**, **Forward Looking function** and **Backward Looking function**. The list of the label types for each of the three main levels is given in Table 2. Table 3 gives breakdown of dialogic labels for the 6241 turns in the corpus.

Information level

In contrast to DAMSL, we do not distinguish between the information level labels for `Task` and `Task-Management`, and we have added a tag to denote the case where the utterance is `Out-of-Task`. It can be seen in Table 4 that over 80% of the utterances are task related and only 2% are obviously `Out-of-Task` (these are typically comments about events in the world, or private conversations). The remaining 17% are annotated with a `Communication` label, signifying that the communication has no direct link with the task (for example during the closing procedure of the dialog). Yes/no responses related to a question about the task are annotated with the `Task` label.

Task	**81%**
C: *mon numéro de compte est le 251* (my account number is 251)	
Out-of-Task	**2%**
C: *moi personnellement je je bricole un peu mais sans plus* (me personally I I play around a bit but not more)	
Communication	**17%**
A: *au revoir* (goodbye)	

Table 4: Proportion and examples of the different **Information level** labels. All sentences have an Information label.

Forward Looking function level

The labels for the Forward Looking function level are shown in Table 5. Most of the `Statement` labeled sentences are assertions (73%). Reassertions (12%) are often due to misunderstandings. For the `Influence-on-listener` values we added a distinction between `Explicit` and `Implicit` information requests (see Table 6). Most of time the information request is explicit (58%), but in about 20% of the cases it is implicit. For example in Table 6 the client simultaneously gives some information to the agent *I can't access my accounts* and implicitly makes a request for an explanation. These kinds of utterances are often observed early on in the dialogs.

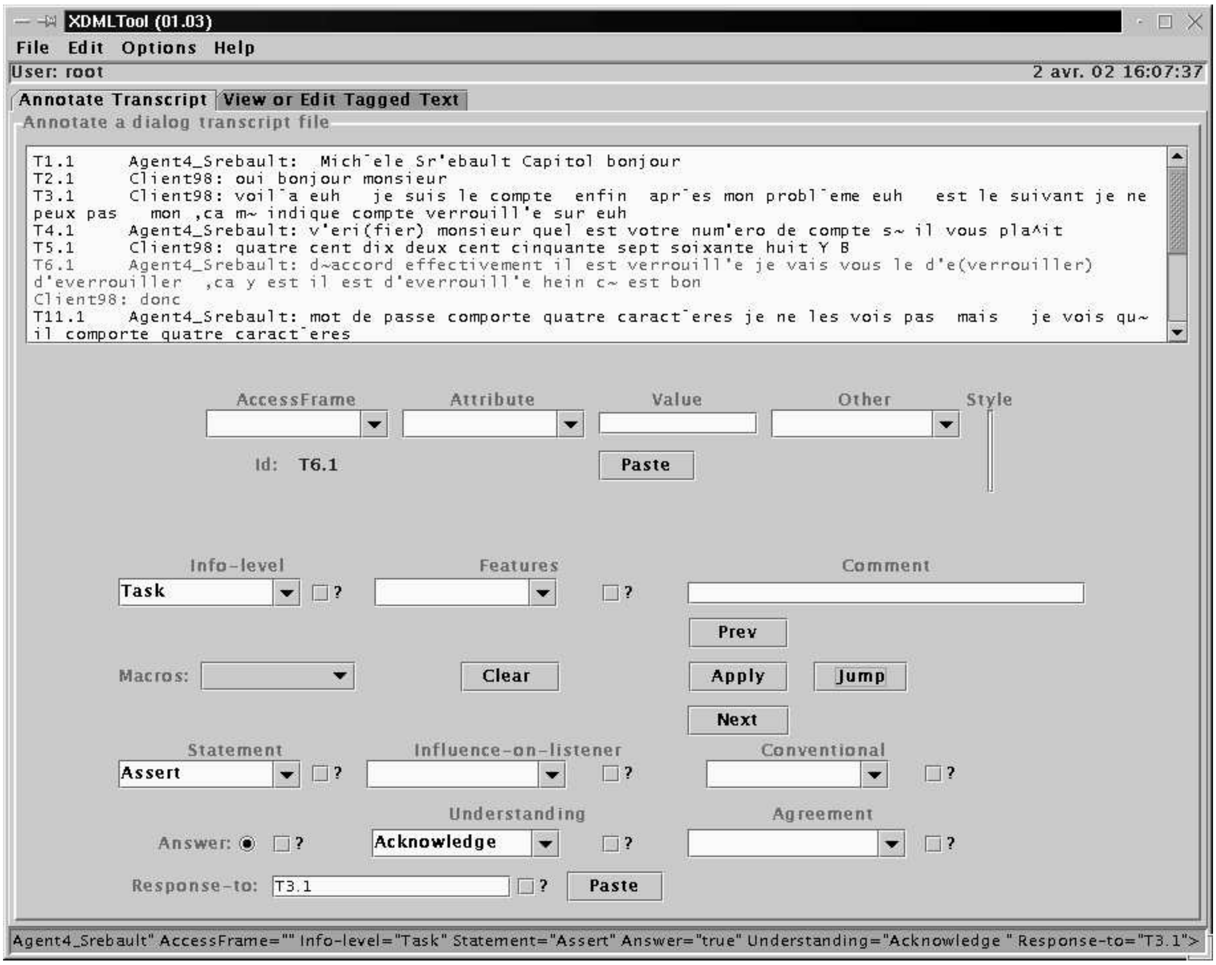
Backward Looking function level

For the Backward Looking function level, the `Understanding` labels (Table 8) concern the actions taken by the speakers in order to ensure that they understand each other as the conversation proceeds. The most frequent understanding label is `Acknowledgment` (70%). Understanding can also be signaled by repeating the information previously given by the other party (`Repetition` 20%), or by making a `Completion` (6%). Only 3% of the Understanding labeled utterances are expressions which indicate a `Non-Understanding`, and 1% are a `Correction`. The `Agreement` labels code the reaction of the speaker to the preceding interlocutor's proposal. The values for the `Agreement` labels in Table 7 show that most of the time speakers agree with their interlocutor (67% if accept and partial accept are counted together). `Rejection`, partial or total, occurs in 18% of the sentences. Rejection does not necessarily imply disagreement in the dialog, it can simply be an answer to a yes/no question. `Agreement` values such as `Maybe`, `I-Don't-Know` and `Exclamation` (around 13% when combined) may indicate that there is a problem in the dialog. The `Answer` label is limited to yes/no values and set to `yes` when the utterance is in response to a previous request. The value of `Response-to` is the number of the utterance that the speaker is reacting to. The `Response-to` may occur immediately after a question, indicating a direct response to the question, or can refer to a question which occurred quite a bit earlier in the dialog.

The different annotation levels can be combined in order to derive some dialog effects. For example as shown in Figure 2 the distance (in terms of number of utterances) between a request (explicit or implicit) and the corresponding answer may be a cue to detect problematic dialogs.

4. Emotion Annotations

Our principal aim in analyzing the emotional behaviors observed in the human-human interactions is to determine which factors may affect human-computer interaction. The usability of an automatic service is certainly dependent upon its ability to adapt its dialog strategy to different user behavior's.



User: root 2 avr. 02 16:07:37

```
T1.1     Agent4_Srebault:  Mich¨ele Sr'ebault Capitol bonjour
T2.1     Client98: oui bonjour monsieur
T3.1     Client98: voil¨a euh   je suis le compte  enfin  apr¨es mon probl¨eme euh   est le suivant je ne
peux pas   mon ,ca m~ indique compte verrouill'e sur euh
T4.1     Agent4_Srebault: v'eri(fier) monsieur quel est votre num'ero de compte s~ il vous pla^it
T5.1     Client98: quatre cent dix deux cent cinquante sept soixante huit Y B
T6.1     Agent4_Srebault: d~accord effectivement il est verrouill'e je vais vous le d'e(verrouiller)
d'everrouiller  ,ca y est il est d'everrouill'e hein c~ est bon
Client98: donc
T11.1    Agent4_Srebault: mot de passe comporte quatre caract¨eres je ne les vois pas  mais   je vois qu~
il comporte quatre caract¨eres
```

AccessFrame: Attribute: Value: Other: Style:

Id: T6.1

Info-level: Task □? Features: □? Comment:

Macros:

Statement: Assert □? Influence-on-listener: □? Conventional: □?

Answer: ● □? Understanding: Acknowledge □? Agreement: □?

Response-to: T3.1 □?

```
Agent4_Srebault" AccessFrame="" Info-level="Task" Statement="Assert" Answer="true" Understanding="Acknowledge " Response-to="T3.1">
```

Figure 1: Screen copy of the XDML tool. Semantic labels (`Accessframe`, `Attribute`, `Value` and `Other`) are not used in this study. The current sentence (`T6.1`, in grey) is annotated as a `Task`-related, `Answer` to `T3.1` which is an implicit request *compte verrouillé* (account locked), as well as an `Assertion` *ca y est il est déverrouillé* (ok, it's unlocked)

Assert	**73%**
C: *mon numéro de compte est le 251* (my account number is 251)	
Reassert	**12%**
C: *mon numéro de compte est le 251*(my account number is 251) A: *pardon* C: *251*	
Offer	**8%**
C: *ben je pourrais envoyer un chèque* (well I could send a check)	
Commit	**5%**
A: *je vais demander à une personne de vous rappeler* (I'll ask someone to call you back)	
Explicit-performative	**2%**
C: *je vais vous je vais vous demander de confirmer* (I'll you I'll ask you to confirm)	

Table 5: Proportion and examples of the different `Statement` labels. The total number of `Statement` labels in the corpus is 2087.

Explicit-Information-Request	**58%**
A: *quel est votre numéro de compte* (what is your account number)	
Implicit-Information-Request	**20%**
C: *oui bonjour euh comme je disais à votre collègue j'ai je ne parviens pas à accéder à mes comptes* (yes hello uh as I was telling your coworker I can't access my accounts)	
Action-Directive	**8%**
C: *j'arrive pas à avoir une personne euh concernant un virement donc euh vous pourriez me passer une personne* (I don't get through to anyone uh concerning a transfer uh can you connect me to someone)	
Please-Wait	**7%**
A: *d'accord ne quittez pas* (ok don't hang up)	
I'm-Listening	**7%**
A: *vente allo* (sales hello)	

Table 6: Proportion of the different `Influence on listener` labels with example sentences. There are a total number 1210 of these labels in the corpus. Utterances with open options are not taken into account.

Accept	**61%**
C: *oui* (yes)	
Accept-Partially	**7%**
C: *oui mais quand même* (yes but anyway)	
Maybe	**2%**
C: *j'ai dû le faire il y a trois ou quatre jours* (I did it about 3 or 4 days ago)	
I-Don't-Know	**6%**
A: *je ne sais pas euh j'en sais rien il faudrait...* (I don't know uh I don't know anything ...)	
Reject	**14%**
A: *ah si ça marche très bien* (but yes, it works very well)	
Reject-Partially	**4%**
C: *non mais ceci dit vous auriez peut-être pu me prevenir* (no well in anycase you could have warned me)	
Exclamation	**6%**
C: *ah bon* (ah yes)	

Table 7: Proportion and examples of the different `Agreement` labels. The total number of `Agreement` labels in the corpus is 895.

Acknowledgment	**70%**
C: *mon numéro de compte est le 251* (my account number is 251) A: *oui* (yes)	
Repetition	**20%**
C: *mon numéro de compte est le 251* (my account number is 251) A: *251*	
Completion	**6%**
A: *alors les coûts euh des coûts de* (so the costs uh the costs of) C: *le coût de SDR* (the cost of SDR)	
Non-Understanding	**3%**
A: *vous acceptez bien les cookies monsieur* (you are sure you accept cookies, sir) C: les ? (the ?) A: *les cookies* (the cookies)	
Correction	**1%**
C: *huit cent quatre vingt quinze* (eight hundred and ninety five) A: *huit cent soixante cinq à la fin* (eight hundred and sixty five at the end)	

Table 8: Proportion and examples of the different `Understanding` labels. There are a total of 1385 `Understanding` labels marked in the corpus.

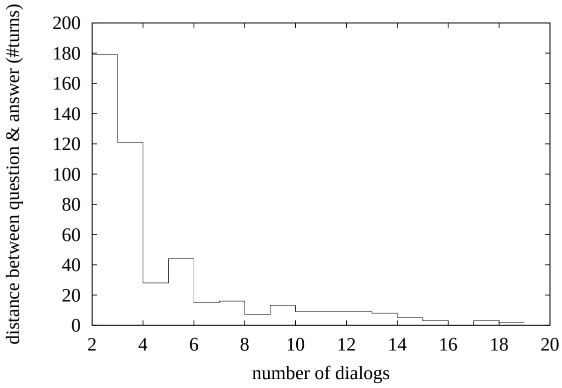


Figure 2: Distance between question and response in dialog turns. The 2286 immediate responses (next dialog turn) are not shown. The portion of the word in parentheses was not pronounced.

Anger	*Fear*	*Satisfaction*	*Excuse*	*Neutral*
5.0%	3.8%	3.4%	1.0%	86.8%

Table 9: Proportion of each emotion label in the dialog corpus labeled by listening to the audio signal.

Choice of labels

Automatic emotion detection is potentially important for customer care in the context of call center services. A task-dependent annotation scheme was developed, keeping in mind that generally the basic affective disposition towards a computer is either trust or irritation. Three of the five classical emotions are retained: *anger* (A), *fear* (F) and *neutral* (N) attitude (the default normal state). In this Web-based stock exchange context, joy and sadness are uncommon emotions and have been excluded from the emotion set. We also considered some of the agent and customer behaviors directly associated with the task in order to capture some of the dialog dynamics. For this purpose, *satisfaction* (S) and *excuse* (E) (apology) were included in the emotion labels. These correspond to a particular class of the speech acts as described in the classical version of pragmatic theory [8]. A 5-point scale was used to annotate the anger and fear emotions. For anger, the levels range from A1 (nervosity) to A5 (aggressivity), whereas for fear the range is from F1 (doubt) to F5 (fear). The remaining annotations are either present or absent: *excuse* (E), *satisfaction* (S), *neutral* (the default) state (N).

Annotation strategy

Two annotators independently listened to the 100 dialogs, labeling each of the sentences with one of the 5 emotions. Sentences with ambiguous labels were judged by a third independent listener in order to decide the final label. Ambiguities occurred on 138 of the 5012 in-task sentences (2.7% of the corpus) and most often involved indecision between neutral state and another emotion: anger (26/138), fear (25/138), and satisfaction (14/138).

It turned out that the annotators did not make use of the full 5-point scale for this task. In most cases, only two levels were used to label the emotions. The highest level marked for anger, A4, was used only 3 times and for fear the highest level F2 was used 16 times. In total, 58 of the 253 sentences labeled with anger were associated with anger levels (A2-A4). Table 9 gives the percentage of sentences in the dialog corpus for each emotion label. For fear and anger, all labels are combined into a single class (F and A, respectively). Based on the auditory classification, sentences with non-neutral labels (F, A, S, E) comprise about 13.2% (649 sentences) of the entire corpus.

5. Dialog Axe Progression Annotations

In developing spoken language dialog systems, we have found a need to represent the progression of the dialog. One of the uses of this dialog axe-based representation is to facilitate the evaluation of the dialog state during the dialog.

Axes for spoken language dialog system

Most task-oriented spoken language dialog systems, such as systems for information retrieval, enable the user to access stored information. Given this view point, we can assess whether an ongoing dialog is running smoothly or is encountering problems. All dialogs evolve, from the first exchange until the end. We can consider that the dialog progression can be represented on two axes: an axe P, which represents the “good” progression of the dialog and an axe A, which represents the accidents which can occur between the system and the user. These axes are represented by respective values, P and A. At each turn, one of the is incremented by 1 (P when all is ok and A when an accident has occured). The number of turns in the dialog is equal to the sum A+P. A third value is used to represent the time (in number of turns) used by the system to repair the accident. The Residual Error (RE) which is incremented when the A value is incremented and decremented when the P value is incremented. This value represents the difference between a perfect (i.e., theoretical) dialog (e.g. without errors, miscommunication...) and the real dialog.

Axes for annotation of human-human dialogs

This kind of representation is not sufficient for uncontrolled human-human dialogs such as those from the AMITIES Stock Exchange call center. In these dialogs we have encountered some turns of speech which are not directly concerned with the task. Moreover, some phenomena, like backchannel acknowledgements, which are helpful for the communication management do not directly contribute to the progression of the dialog with an A or a P value, insofar as we consider the task. Thus, we decided to add two new values for Out-of-Task and Backchannel utterances. The P, A and ER values are unchanged (not incremented or decremented) when the turn is not directly concerned with the task and when it is only a backchannel (except when a backchannel marks a repair). Thus the 5 values are used to annotate this corpus are: T (turn of speech), P (progression), A (accident), ER (residual error) and OT (out of task). The annotation is marked only on the agent’s turn.

The 100 agent-client dialogs have been annotated with the dialog axes values. The entire corpus contains 1136 progressions. There are a total of 252 accidents, which occured in 70 dialogs. 35 of these dialogs have unrepaired accidents at the end of the dialog, with a summed total of 88 residual errors at the end of the dialogs.

The Figure 3 shows a extract of one of the human-human dialogs with the progression axe labels. The full dialog is much longer. Figure 4 plots the Residual Error as a function

A: *d'accord* (ok) [9 2 0 0 1]

C: *bon et euh ça c'est une première chose deuxième chose je j'ai fait des opérations le vingt trois et le vingt cinq* (that's the first thing second thing I did some transactions the 23rd and 25th)

A: *vingt trois et vingt cinq* (23 and 25) [10 2 0 0 1]

C: *oui euh et euh euh* (yes uh uh)

A: *hein* (huh) [11 2 1 1 1]

C: *oui oui et les c'est pour ça que j'ai attendu votre appel je préférais plutôt que d'en discuter et euh les les taux pratiqués ne sont pas du tout ceux euh qui sont pratiqués habituellement hein donc il y a il y a des erreurs euh* (yes yes and that's why I was waiting for your call I would prefer rather than to discuss uh the the rates used are not those generally used uhm therefore there are there are some errors uh)

A: *qui a été prélevé peut-être non* (which was charged maybe or not) [12 3 1 0 1]

C: oh je ne je ne sais pas on puis il y a des moments où il y a eu encore des problèmes informatiques euh donc euh or que j'avais eu à ce moment-là m'a dit que euh il allait resignaler où vous aviez changé de ... (oh I I don't know and sometimes there are some technical problems uh uh if I knew at this time ...)

A: *c'est deux opérations* (it is 2 transactions) [13 3 2 1 1]

C: *plusieurs hein il y en a plusieurs j'en ai fait quatre ou cinq à peu le mercredi vingt trois et le vendredi vingt cinq* (several uhm there are several that I did 4 or 5 Wednesday the 23rd and Friday the 25th)

A: *ok donc ca je le note* (ok I'm writing them) [14 4 2 1 1]

Figure 3: Example of dialog progression annotation with the 5 axe values [T,P,A,RE,OT]. T: Turn number, P: Progression, A: Accident, RE: Residual Error, and OT: out-of-task. Turns 9-14 in Figure 4. Two uncorrected RE remain at turn 14.

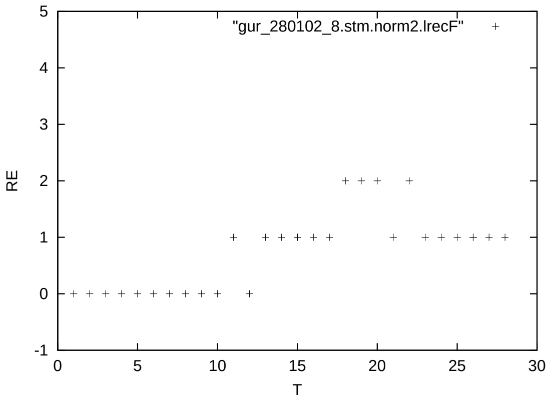


Figure 4: Residual error vs dialog turn in an Agent-Client dialog.

of the dialog turns, giving a graphical representation of the dialog progression. It can be noticed that several turns may be needed to get a dialog back on track and not all errors are corrected.

6. Predictive Functions and Factor Analysis

The three annotation types described above are used to determine a set of factors with which predictive models are estimated.

Choice of parameters

For this experiment, 14 parameters were extracted from the annotated corpus. These parameters primarily denote negative factors in the 3 annotation types (Non-Understanding, Accident, Anger,...) which can be expected to have an influence on the dialog quality. Five parameters are taken from the Dialogic annotations: at the `Statement` level, Reassert (REA); at the `Agreement` level, Reject (REJ) and I-Don't-Know (IDK); and at the `Understanding` level, Non-Understanding (NUN) and correct (COR). Three parameters concern the dialog axe progression: Residual Error (RER), Accident (ACC) and Progression (PRO). The five emotion labels are kept: Fear (FEA), Anger (ANG), Neutral state (NEU), Excuse (EXC) and Satisfaction (SAT). The last parameter is the dialog length (LEN). Some of these parameters can be categorized as utterance-level features (emotion and dialogic labels), and some others are per-turn features (dialog axe progression parameters). As a first measure of the dialog quality a global predictive feature vector is extracted for each dialog. This vector is formed by summing and normalizing all of the occurrences of each of the selected 14 parameters.

Methodology and analysis

Table 10 shows the correlations between the 14 parameters. Correlations higher than 0.4 are shown in bold. There are very high correlations between dialog length and dialog progression with neutral state, which is to be expected since over 86% of the sentences have this label. Another notable correlation is between Residual Error and Accident, which is also expected.

We used classical multiple linear regression techniques to find which combination of factors are able to predict parameters such as Accident and Residual Error or emotion (Anger and Fear) in a dialog. Different multiple regression models were estimated by adding and dropping terms as appropriate using ANOVA.

Table 11 shows some prediction models for detecting dialogs with problems, in particular for Accidents and Residual Errors. A correct prediction for the parameter ACC (p=0.0) is obtained with the predictive factors: ERR, ANG, EXC, FEA, COR and REJ (first entry). Taken together these factors explain 81.6% of the variance of accidents, with the highest contribution from RER. The next 3 models remove the RER factor, which is highly correlated with accidents and may mask the contributions of other factors. The second entry explains 65.5% of the variance of the accidents. Comparing the 3rd and 4th entries, the Emotion factors EXC, FEA and ANG seem to be better predictors of accidents (58.8%) than the dialogic factors (47.6%) retained here. It can be inferred that the Emotion factors account for most of the explanation of the 2nd model.

Models were also built to predict the RER at the end of the dialog, which is an important indication of the overall dialog success. The first model is able to explain 44.6% of

	ACC	RER	PRO	FEA	ANG	SAT	EXC	NEU	IDK	COR	NUN	REA	REJ	LEN
ACC	1.0													
RER	**0.81**	1.0												
PRO	**0.52**	0.31	1.0											
FEA	**0.46**	**0.41**	**0.47**	1.0										
ANG	**0.63**	**0.54**	**0.41**	0.30	1.0									
SAT	0.15	0.07	0.3	**0.42**	0.12	1.0								
EXC	**0.51**	0.35	0.21	0.10	0.32	0.04	1.0							
NEU	0.35	0.17	**0.84**	0.29	0.30	0.21	0.10	1.0						
IDK	0.38	0.38	0.38	0.38	0.30	0.1	0.08	0.11	1.0					
COR	0.07	0.01	0.17	-0.06	-0.15	0.02	-0.05	0.04	-0.03	1.0				
NUN	0.27	0.28	0.17	0.15	0.21	0.01	0.12	0.08	0.07	0.10	1.0			
REA	**0.56**	**0.43**	**0.56**	0.34	**0.50**	0.19	0.27	0.36	0.21	0.01	0.29	1.0		
REJ	**0.54**	0.35	0.39	**0.40**	**0.45**	0.27	0.25	0.23	0.16	-0.03	0.05	**0.59**	1.0	
LEN	0.33	0.16	**0.82**	0.29	0.30	0.23	0.1	**0.99**	0.11	0.04	0.09	0.34	0.22	1.0

Table 10: Correlations among the 14 selected factors. ACC: accident, RER: residual error, PRO: progression, FEA: fear, ANG: anger, SAT: satisfaction, EXC: excuse, NEU: neutral, COR: correct, NUN: non-understanding, REA: reassert, REJ: reject, and LEN: dialog size.

Variable	*Main Predictors*						*Explanation*
ACC	.55 · RER	.22 · EXC	.18 · REJ	.17 · ANG	.12 · COR	.10 · FEA	81.6%
ACC	.34 · ANG	.33 · EXC	.22 · FEA	.20 · REJ	.17 · COR	.12 · IDK	65.5%
ACC	.42 · ANG	.35 · EXC	.32 · FEA				58.8%
ACC	.34 · REJ	.27 · IDK	.25 · REA	.16 · NUN			47.6%
RER	.28 · ANG	.20 · FEA	.18 · EXC	.14 · IDK	.13 · NUN	.10 · REA	44.6%
RER	.38 · ANG	.29 · FEA	.19 · EXC				39.9%
RER	.29 · IDK	.19 · REA	.19 · NUN	.17 · REJ			31.9%
ANG	.33 · ACC	.21 · REA	–.18 · COR	.14 · IDK	.07 · RER		48.6%
FEA	.24 · IDK	.22 · ACC	.21 · REJ				30.6%

Table 11: Prediction models for **ACC**, **RER**, **ANG**, **FEA**. The weighted main factors predict the variable with the percentage given in the Explanation column.

the variance of the residual dialog progression errors with a p_value of 4.496e-10. Anger is also seen to be correlated with error at the end of the dialog and is a good predictor of dialog problems.

Finally, we tried to predict emotions such as Anger and Fear. Client Anger can be partially explained with dialog axe progression accidents, and dialogic labels (reassertion, correction), but Fear is unable to be predicted with better than 30% using any combination of these 14 parameters. Client anger is to some degree correlated with the need to repeat information, but the negative weight of correction seems to imply that correcting errors is not a big deal. Problems arise when the one of the interlocuters is unable to correct an error. These first experiments will be validated on a substantially larger corpus.

7. Conclusions

The present study reports recent developments in annotating a corpus of human-human dialogs for a Web-based Stock Exchange call center. This work is carried out in the context of the AMITIES project which aims to explore novel technologies for adaptable multilingual spoken dialog systems. Central to the project is the study and modelization of large corpora of human-human and human-computer dialogs which serve as the basis for system development.

We have annotated a initial corpus of 100 dialogs with three types of annotations: dialogic labels (DAMSL-style), dialog progression axe labels, and emotion labels. The annotations were carried out independently so as to minimize any biases. Using standard multiple linear regression techniques, a predictive function of dialog problems was derived, estimating the relative contributions of various factors extracted from dialogic, progression and emotion annotations. These measures are able to explain about 80% of the dialog accidents. The observed correlations between DAMSL-like dialogic labels and the annotations for emotion and dialog axe progressions provide evidence that these latter annotation types are relevant.

Over the next year, the annotations will be extended to over 1000 dialogs. In addition to the annotations used in this work, we have started to define semantic level annotations which concern the focus of the interaction and the attribute/value representation.

This data will also be used to build a number of different models to detect and predict utterance topics, emotions and dialog acts based on different sources of evidence: lexical, semantic, emotion and discourse sequence. The relationships between dialog annotations and meta-annotations will be used to determine features which can be automatically extracted in order to dynamically adapt the dialog strategy of the spoken language dialog system accordingly.

8. Acknowledgments

Thanks are due to Catherine Mathon and Isabelle Wilheim for helping with the annotations of the dialogue corpus.

REFERENCES

[1] *1st SIGDIAL workshop on Discourse and Dialog*, http://www.sigdial.org/sigdialworkshop/program.html

[2] http://www.dcs.shef.ac.uk/nlp/amities

[3] J. Allen and M. Core, "Draft of DAMSL: Dialog Act Markup in Several Layers," October 1997. http://www.cs.rochester.edu/research/trains/annotation

[4] H. Bonneau-Maynard, L. Devillers, S. Rosset, "Predictive performance of dialog systems", *LREC*, 2000.

[5] C.M. Lee, S. Narayanan, R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal", *ASRU*, 2001.

[6] D. Luzzati, "Recherches sur le dialogue homme-machine: mod`eles linguistiques et traitements automatiques", Th`ese d'état, Paris III, 1989.

[7] V. Petrushi, "Emotion in speech: recognition and application to call centers", *Artificial Neural Network ANNIE'99*.

[8] JR. Searle, D. Vanderveken, "Fondations of Illocutioanry Logic", Cambridge: CUP, 1985.

[9] M. Walker and D. Litman and C. Kamm and A. Abella, "Paradise: a general framework for evaluating spoken dialog agents", ACL/EACL, 1997.

[10] M. Walker and R. Passonneau, "DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems," *Human Language Technology Conference*, San Diego, March, 2001.