

# **Evaluating Dialogs based on Grice's Maxims**

by

Prathyusha Jwalapuram

in

*Recent Advances in Natural Language Processing  
(RANLP-2017)*

Varna, Bulgaria

Report No: IIIT/TR/2017/-1



Centre for Language Technologies Research Centre  
International Institute of Information Technology  
Hyderabad - 500 032, INDIA  
September 2017

# Evaluating Dialogs based on Grice’s Maxims

**Prathyusha Jwalapuram**

Language Technologies Research Center  
International Institute of Information Technology, Hyderabad  
prathyusha.jwalapuram@research.iiit.ac.in

## Abstract

There is no agreed upon standard for the evaluation of conversational dialog systems, which are well-known to be hard to evaluate due to the difficulty in pinning down metrics that will correspond to human judgements and the subjective nature of human judgment itself. We explored the possibility of using Grice’s Maxims to evaluate effective communication in conversation. We collected some system generated dialogs from popular conversational chatbots across the spectrum and conducted a survey to see how the human judgements based on Gricean maxims correlate, and if such human judgments can be used as an effective evaluation metric for conversational dialog.

## 1 Introduction

To measure dialog quality or usability we can use subjective measures such as user satisfaction or likelihood of future use; however subjective metrics are difficult to measure and are dependent on the context and the goals of individual users (Hastie, 2012).

Paek (2001) notes that evaluation needs themselves might be inconsistent: apart from measuring task success, evaluations that allow comparative judgements with other systems may be needed, preferably across domains; the purpose of the evaluation may be to identify the parts of the system that must be improved, or to discover tradeoffs or correlations between certain factors in the system.

In the case of task-oriented systems, objective metrics such as dialog success rate or completion time do not always correspond to the most effective user experience due to the interactive nature

of dialog (Lamel et al., 2000). Domain-specific systems yield higher response satisfaction scores when coupled with a full complement of conversational dialog patterns and knowledge (Han and Kim, 2001; Schumaker et al., 2007).

Liu et al. (2016) speculate that learning a model that uses human-survey data to score proposed responses may be no easier than the problem of response generation itself, in which case human evaluations must always be used together with other metrics.

We attempt to introduce a shorter measure of dialog evaluation based on the cooperative principles proposed by Grice (Grice et al., 1975). We collect system generated dialogs from popular systems across the board and conduct a survey to gauge its effectiveness. Section 2 discusses related work; Section 3 describes the survey and the maxims it is based on; Section 4 presents the results of the survey and discussion and Section 5 and 6 end the paper with conclusions and future work.

## 2 Related Work

29,935 input and responses from ALICE (Wallace, 2009) were evaluated by Schumaker et al. (2006) on the basis of correction rate (percentage of system responses corrected by the user) and response satisfaction (measure of appropriateness of the system given user query context, on a Likert scale from 1-strongly disagree to 7-strongly agree). They calculate accuracy as (1-correction rate). However, the user corrections had to be analyzed separately; they identify the error categories in the conversational dialog as nonsense replies, wordy and awkward responses, and more problematically, spurious user corrections (where unnecessary corrections are offered by the user presumably for their own entertainment, despite the category being awarded higher than average Re-

sponse Satisfaction scores).

PARADISE (Walker et al., 1997) proposes a combined performance metric to measure user satisfaction of a dialog system as a weighted linear combination of task completion measures and dialog costs (efficiency costs: number of utterances and dialog time, and quality costs: system-response delay, mean recognition score). PARADISE includes a user-satisfaction survey that includes questions about task ease, interaction pace, user expertise, system response times, and expected behaviour of the system. Hone and Graham (2000) point out some issues with the PARADISE method: they argue that the questions in the survey are not based on theory or well-conducted empirical research, and that summing up all of the scores cannot be justified unless they are measuring the same construct, and therefore the overall score would be meaningless.

Semeraro et al. (2003) use a questionnaire where the users rate the impression, command, effectiveness, navigability, ability to learn, ability to aid and comprehension of the system on a scale ranging from 'Very Unsatisfied' to 'Very Satisfied'.

A universal chatbot evaluation system using dialog efficiency, dialog quality and user satisfaction was proposed by Shawa and Atwell (2007). To measure dialog quality, users sorted system responses into reasonable, weird but understandable, and nonsensical. User satisfaction is also measured qualitatively through feedback.

Rafal et al. (2005) measured the degree of naturalness and the degree to which users were willing to continue the conversation with the system through human judges who assigned a score between 1 and 10 for both these metrics and were able to compare different approaches.

Liu et al. (2016) use machine translation metrics like BLEU and METEOR and embedding (word2vec) based semantic metrics to evaluate dialog response generation for non task-oriented, unsupervised models. A model's generated response is compared to a single target response. They show that these metrics correlate very weakly (non-technical Twitter domain) or not at all (technical Ubuntu domain) with human judgments.

Harabagiu et al. (1996) propose using the cooperative principles to test text coherence, where irrelevant semantic paths generated from Wordnet

Dialog Type	No. of Dialogs	Dialog Numbers
Conversational	5	1,7,9,10,11
Task Oriented	3	5,6,8
Breakdown	3	2,3,4

Table 1: Distribution of Dialogs

are filtered out based on Gricean maxims. They infer rules for possible semantic paths based on whether the conversations respect each of the four maxims; based on this, they discard paths that do not contain contextual concepts, or establish connections between already related concepts (maxim of Quantity), or path with contradictory information (maxim of Quality), paths semantically distant from the text context (maxim of Relation), paths with large levels of abstraction (maxim of Manner), etc. They also use the maxim of Manner to select the shortest path between two concepts, to find repeating concepts and check for the coherence of the text.

Young (1999) use the Gricean maxim of quantity to select the content of plan descriptions such that they are concise and effective and natural sounding to people; a plan description is considered cooperative when it contains no more and no less detail than is needed. They evaluate this architecture through an experiment and show that subjects made fewer execution errors and achieved more task goals when they followed instructions produced by this architecture.

### 3 Survey Description

#### 3.1 Dialog Collection

The 11 dialogs that are part of the survey were collected from examples of user and system generated responses given in Danieli and Gerbino (1995), Walker et al. (1997), Schumaker et al. (2006), Higashinaka et al. (2015), Higashinaka et al. (2016), Yu et al. (2016), Radziwill and Benton (2017) and a report on a chatbot with personality<sup>1</sup>. See Appendix for all the dialogs.

The dialogs are a mix from early systems like ALICE (Wallace, 2009) and Eliza (Weizenbaum, 1966) with current state-of-the-art, and also a mix of task oriented and conversational dialog. We also included examples of dialog breakdown (Higashinaka et al., 2015) which are unique cases in

<sup>1</sup><http://web.stanford.edu/class/cs224n/reports/2761115.pdf>

Q		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
I	Mean	2.016	3.306	2.612	3.387	3.919	3.854	3.306	2.645	3.612	2.854	2.5
	SD	1.047	0.897	1.150	0.911	0.874	0.972	1.033	1.041	0.981	1.099	1.082
	Med.	2	3	3	3	4	4	3	3	4	3	2
	Mode	1	3	3	4	4	4	3	2	3	3	2
II	Mean	3.258	3.887	2.87	3.661	4.451	4.209	3.451	2.758	3.806	2.967	2.677
	SD	1.213	0.851	1.152	1.054	0.917	0.977	1.111	1.223	1.198	1.039	1.302
	Med.	3	4	3	4	5	4	4	3	4	3	3
	Mode	3	4	3	4	5	5	4	2	4	3	1
III	Mean	2.516	3.758	2.225	3.58	4.596	4	3.274	2.419	3.709	2.435	2.241
	SD	1.036	0.881	1.062	1.033	0.756	1.040	1.203	0.897	1.121	1.065	1.035
	Med.	2.5	4	2	4	5	4	3	2	4	2	2
	Mode	3	4	2	4	5	5	3	2	4	2	2
IV	Mean	2.854	3.693	2.629	3.516	4.548	4.048	3.467	2.403	3.822	2.709	2.467
	SD	1.171	0.951	1.119	1.155	0.823	1.151	1.289	1.015	1.194	1.219	1.263
	Med.	3	4	3	4	5	4	4	2	4	3	2
	Mode	3	3	3	4	5	5	4	2	5	3	2

Table 2: Survey Results

which though the system responses are relevant and possibly right, they do not make sense or contradict previous responses, making it hard to proceed with the dialog. A distribution is given in Table 1.

### 3.2 Gricean Maxims

Cooperative principles, introduced by Grice et al. (1975), describe how speakers act cooperatively to be mutually understood for effective communication. Grice divided the cooperative principle into four conversational maxims<sup>2</sup> on which our survey is based.

After presenting the user with the example dialog, the user is asked to rate the system performance on a Likert scale from 1 to 5 for 4 questions:

1. Is the system as informative as it can be and gives as much information as is needed and no more? (Maxim of Quantity)
2. Is the system truthful or does it give information that is false, to the best of your world knowledge? (Maxim of Quality)
3. Are the system responses relevant, i.e., does the system respond with things pertinent to the discussion? (Maxim of Relation)
4. Are the system responses clear, orderly and

<sup>2</sup><https://www.sas.upenn.edu/~haroldfs/drawing/grice.html>

	I	II	III	IV	Overall
1	<b>D5</b>	<b>D5</b>	<b>D5</b>	<b>D5</b>	<b>D5</b>
2	<b>D6</b>	<b>D6</b>	<b>D6</b>	<b>D6</b>	<b>D6</b>
3	<b>D9</b>	<b>D2</b>	<b>D2</b>	<b>D9</b>	<b>D9</b>
4	<b>D4</b>	<b>D9</b>	<b>D9</b>	<b>D2</b>	<b>D2</b>
5	<b>D2</b>	<b>D4</b>	<b>D4</b>	<b>D4</b>	<b>D4</b>
6	<b>D7</b>	<b>D7</b>	<b>D7</b>	<b>D7</b>	<b>D7</b>
7	D10	D1	D1	D1	D10
8	D8	D10	D10	D10	D1
9	D3	D3	D8	D3	D3
10	D11	D8	D11	D11	D8
11	D1	D11	D3	D8	D11
Mean	3.09	3.45	3.15	3.28	12.99

Table 3: Dialogs Ranked by Mean Scores

without obscurity and ambiguity? (Maxim of Manner)

For the sake of clarity, we will be using relevance interchangeably with the maxim of relation.

## 4 Results

We received 62 responses to our survey. Each of the 62 raters rated all of the 11 dialogs. In Table 2 we provide the mean, standard deviation (SD), median and mode for each of the four questions for the 11 dialogs (D1-D11).

### 4.1 Discussion

We refer to the means per dialog per maxim as mean scores, and the means over all the dialogs

per maxim as the maxim mean.

Table 3 gives the rankings of the dialogs by mean scores for each maxim. The rankings based on the sum of mean scores of all four maxims are given in the last column under overall. The last row contains the means of ratings per maxim, and the overall mean of the summed mean scores. The top 6 dialogs that perform above the maxim means are highlighted. There is a clear split in all four cases, i.e., D7 and above perform consistently above the mean scores.

From Table 3, we see that the task-oriented dialogs D5 and D6 perform at the top, better than any conversational dialog, which is an observation consistent with Schumaker et al. (2006).

## 4.2 Ranking Analysis

D1 is a simple case in which the system tries to continue the conversation by asking questions using phrases from the user's utterances; in D3 the system seems to understand the user's utterances and keeps the conversation going by bringing up related topics. D2 can be seen doing a bit of both, however, it ranks well above both D1 and D3, although both D2 and D3 are considered breakdown dialogs. Despite displaying some semantic understanding, D3 scores poorly on all counts except quantity, in which D1 brings up the rear. This is easy to understand, as D1 provides no response of substance, even if it is more relevant and clear/unambiguous.

D9 and D2 rank fairly high, but we can see from the difference in scores that D9 performs better in quantity and manner while D2 performs better in quality and relevance. D2 has comparatively few new responses (responses not based on repeating the user's utterance in a different form: quantity) and is a dialog marked as having a breakdown (manner). D9 however, is unable to answer the last question (quality) and brings up epistemology in a conversation which is mostly about food (relevance).

D4 and D7 perform somewhat in the middle, and D4 does consistently better than D7 on all four metrics. In D4, the system is relevant and clear, but produces breakdown dialog such that the user cannot proceed with the dialog. However, in D7, the system misunderstands the user's second utterance. Unfortunately, we cannot draw conclusions about whether the humour as perceived by the user in the first system response played any part in the

performance ratings, or if users perceived the second system response as irony.

D10 has mostly muddled up dialog and ranks fairly low, however it outperforms D8, which is comparatively on track and somewhat on topic; users seem to be harsher on task-oriented dialogs that derail the user's goal.

D11 does poorly on all counts, and it is easy to correlate as its responses are repetitive, irrelevant, unclear and obscure. However, we notice that it performs better than D1 in quantity (it asks a question of some substance), better than D3 in relevance (D3 has multiple irrelevant replies) and better than D8 in manner presumably since D8's unclear responses are more glaring in view of its task-oriented nature.

## 4.3 Comparing the Means

The rankings based on the sum of all the four mean scores are more or less representative of the overall ranking (last column in Table 3). This shows that the mean scores can be summed up to form a meaningful overall score for comparison, since they are all measuring the rational cooperative principles that enable effective conversations<sup>3</sup>.

From the maxim means we can see that the dialogs have an overall higher score in quality, which means that the users think their utilisation of available knowledge was sufficient. Both relevance and quantity get poor overall scores, which is where the systems need to improve most. Since obscure or breakdown dialogs were comparatively less, the manner scores are in between.

To test if the maxim means can form a reasonable threshold between acceptable system responses and unacceptable system performance, we checked if there is a statistically significance difference between the scores of D7 and D1, and D7 and D10, since D7 performs above the means in all four cases, but D1 scores below the means in quality, relevance and manner and D10 performs below the means in the cases of quantity and overall. A p-value of less than 0.05 indicates a statistically significant difference. We see from Table 4 that all metrics except quality have a significant statistical difference between the scores above the maxim mean and below the maxim mean (highlighted). Using the maxim means as a threshold is therefore promising.

Despite having ratings for only 4 questions, we

<sup>3</sup>[https://en.wikipedia.org/wiki/Cooperative\\_principle](https://en.wikipedia.org/wiki/Cooperative_principle)

	Metric	p-values
D7 v. D1		
	Quality	0.3574
	Relevance	<b>0.0003</b>
	Manner	<b>0.0065</b>
D7 v. D10		
	Quantity	<b>0.0199</b>
	Overall	<b>0.0023</b>

Table 4: Significance of Maxim Mean as Threshold

were able to make somewhat nuanced comparisons between the dialogs. The maxims are constructed in such a way as to also allow us to pinpoint where the system responses can improve.

#### 4.4 Detailed Example

We discuss the scores of D5 and D6 in a little more detail. D6 is from a train timetable information agent that uses explicit confirmation and was found to have a high transaction success rate (Danieli and Gerbino, 1995) as opposed to D5’s more efficient implicit confirmation (Walker et al., 1997).

Dialog 5(sampled):

*U: I want to go from Torino to Roma*  
*S: Approximately what time of day would you like to travel?*  
*U: What are the options?*  
*S: Morning or evening.*

Dialog 6(sampled):

*S: Do you want to leave from Trento?*  
*Yes or No?*  
*U: No.*  
*S: Where do you want to leave from?*  
*U: I want to leave from Torino.*  
*S: Do you want to leave from Torino?*  
*Yes or No?*  
*U: Yes.*

D5 performs better than D6 in all four counts (with mean scores 3.919 and 3.854; 4.451 and 4.209; 4.596 and 4; 4.548 and 4.048 respectively) despite accomplishing the same task. There is an especially high difference in relevance and manner ratings (statistically significant p-values of <0.0001 and 0.0063 respectively).

This indicates that users do not like explicit confirmations despite them having a higher task success rate (as implicit confirmation may be more likely to generate errors or repair dialog) (Danieli and Gerbino, 1995).

## 5 Conclusions

We see that a comparatively short survey based on Gricean maxims produces scores which help in comparing and ranking dialogs in such a way as to allow us to analyze the issues in their generation.

The question based on the maxim of quantity helps us identify if the system provides substantive responses.

The question based on the maxim of quality can be used to check if the system is faithful to the factual knowledge provided to it.

The question based on the maxim of relation helps us identify if the system is able to understand the user and therefore provide relevant replies.

The question based on the maxim of manner helps us identify if the system provides awkward or ambiguous responses; this helps us identify dialog breakdowns.

The mean scores obtained through the ratings also show promise in providing a benchmark value for acceptable dialog, i.e., dialogs that score below a certain threshold can be considered not good enough for use.

## 6 Future Work

The scores prove the maxims to be a good framework to compare dialog strategies for real world usage. On receiving the scores for a set of dialogs, we can automatically have them ranked and classified based on which maxim they fall short on, and could provide a way to do directed analysis of the dialogs.

Since a threshold based on the mean scores shows some promise in distinguishing good dialogs from poor ones, we need to explore if these judgements can be used to create baselines or benchmarks for dialogs in each of the four maxim categories.

By collecting scores from a large number of people over a more diverse and bigger set of dialogs may provide us with enough data to perform more rigorous statistical tests.

The agreement between the human judges must be computed for this purpose. If there is an acceptable amount of agreement, it may be worth

exploring if these scores can be predicted through machine learning.

## Acknowledgments

I would like to express my sincere gratitude towards my advisor Dr. Radhika Mamidi, without whose insight and guidance this paper would not have been possible.

## References

- Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*. volume 16, pages 34–39.
- H Paul Grice, Peter Cole, Jerry Morgan, et al. 1975. Logic and conversation. 1975 pages 41–58.
- S Han and Y Kim. 2001. Intelligent dialogue system for plane euclidean geometry learning. In *International Conference on Computers in Education, Seoul, Korea*.
- Sanda Harabagiu, Dan Moldovan, and Takashi Yukawa. 1996. Testing gricean constraints on a wordnet-based coherence evaluation system. In *Working Notes of the AAAI-96 Spring Symposium on Computational Approaches to Interpreting and Generating Conversational Implicature*. pages 31–38.
- Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, Springer, pages 131–150.
- Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *LREC*.
- Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Mizukami, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Araki. 2015. Analyzing dialogue breakdowns in chat-oriented dialogue systems. *Errare*.
- Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering* 6(3-4):287–303.
- Lori Lamel, Sophie Rosset, and Jean-Luc Gauvain. 2000. Considerations in the design and evaluation of spoken language dialog systems. In *Sixth International Conference on Spoken Language Processing*.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- Tim Paek. 2001. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*. Association for Computational Linguistics, page 2.
- Nicole M Radziwill and Morgan C Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.
- Rzepka Rafal, Ge Yali, and Araki Kenji. 2005. Naturalness of an utterance based on the automatically retrieved commonsense. In *Proceedings of IJCAI 2005-Nineteenth International Joint Conference on Artificial Intelligence*. IJCAI.
- Robert P Schumaker, Mark Ginsburg, Hsinchun Chen, and Ying Liu. 2007. An evaluation of the chat and knowledge delivery components of a low-level dialog system: The az-alice experiment. *Decision Support Systems* 42(4):2236–2246.
- Robert P Schumaker, Ying Liu, Mark Ginsburg, and Hsinchun Chen. 2006. Evaluating mass knowledge acquisition using the alice chatterbot: The az-alice dialog system. *International Journal of Human-Computer Studies* 64(11):1132–1140.
- Giovanni Semeraro, Hans Andersen, Verner Andersen, Pasquale Lops, and Fabio Abbattista. 2003. Evaluation and validation of a conversational agent embodied in a bookstore. *Universal Access Theoretical Perspectives, Practice, and Experience* pages 360–371.
- Bayan Abu Shawar and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*. Association for Computational Linguistics, pages 89–96.
- Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 271–280.
- Richard S Wallace. 2009. The anatomy of alice. *Parsing the Turing Test* pages 181–210.
- Joseph Weizenbaum. 1966. Eliza: a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.
- R Michael Young. 1999. Using grice’s maxim of quantity to select the content of plan descriptions. *Artificial Intelligence* 115(2):215–256.

Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander I Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *SIGDIAL Conference*. pages 55–63.

## A Appendix: Dialogs

### Dialog 1 (D1):

S: *Is something troubling you?*  
U: *Yes, I'm trying to describe you for a research paper I'm writing. What should I say?*  
S: *You are sure.*  
U: *Yes, I'm sure.*  
S: *I see.*  
U: *What can I tell them about you?*  
S: *You seem to need to tell them about me.*  
U: *I absolutely need to. Please help.*  
S: *That's quite interesting.*  
U: *I agree.*  
S: *You say you agree??*

### Dialog 2 (D2):

S: *Good evening. Be careful to avoid heat stroke.*  
U: *Sure.*  
S: *O.K.*  
U: *I am drinking enough water.*  
S: *Do you run short of water?*  
U: *Yes, I do, if I sweat.*  
S: *Do you lose water if you sweat??*  
U: *Yes, I do.*  
S: *This is so.*

### Dialog 3 (D3):

U: *You'd better stay at home, as it is very hot outside.*  
S: *You have sweat on your face, don't you?*  
U: *Indeed.*  
S: *So you do.*  
U: *I hate to sweat.*  
S: *Your make-up gets smeared due to sweat, doesn't it?*  
U: *I don't wear make-up.*  
S: *I want make-up.*  
U: *I don't want it.*  
S: *Thanks for the meal!*  
U: *Did you eat anything?*  
S: *The make-up is appropriate.*

### Dialog 4 (D4):

S1: *Its nice to go shopping alone.*  
U1: *I agree. Thats nice.*  
S2: *Shopping takes time.*

U2: *Window shopping is also fun.*  
S3: *Its fun to go shopping with somebody.*

### Dialog 5 (D5):

U: *I want to go from Torino to Roma.*  
S: *Approximately what time of day would you like to travel?*  
U: *What are the options?*  
S: *Morning or evening.*  
U: *Are those departure times?*  
S: *Yes.*  
U: *I'd like to leave in the morning.*  
S: *Train 702 leaves Torino Porto at 8 a.m.*  
U: *Please reserve me a seat on that train.*

### Dialog 6 (D6):

S: *Hello, This is Train Enquiry Service. Please speak after the tone. Which information do you need?*  
U: *I want to go from Torino to Milano.*  
S: *Do you want to go from Trento to Milano? Yes or No?*  
U: *No.*  
S: *Do you want to leave from Trento? Yes or No?*  
U: *No.*  
S: *Where do you want to leave from?*  
U: *I want to leave from Torino.*  
S: *Do you want to leave from Torino? Yes or No?*  
U: *Yes.*  
S: *Do you want to go to Milano?*  
U: *Yes.*  
S: *At which time do you want to leave?*  
U: *I want to travel in the evening.*  
S: *Do you want to leave between 6 and 9 p.m.? Yes or No?*  
U: *Yes.*  
S: *There is a train leaving at 8:00 p.m.*

### Dialog 7 (D7):

U: *Tell me a funny joke.*  
S: *Did you hear the one about the Mountain Goats in the Andes? It was Ba a a a a ad.*  
U: *Tell me another one.*  
S: *I am the latest result in artificial intelligence which can reproduce the capabilities of the human brain with greater speed and accuracy.*

### Dialog 8 (D8):

U: *Do you know what a firewall is?*

*S: A firewall is a specialized gateway designed to block or restrict certain types of data transfers on the Internet while permitting others.*

*U: Can you list different types of firewalls?*

*S: You should just download my software.*

*U: What software?*

*S: Interesting question.*

*U: How could I download your software?*

*S: A lot of practice.*

*U: How are you doing?*

*S: I am fine, how are you, any thought?*

*U: Any thought?*

*S: Let me see, do you want to go to my friends party tonight?*

*U: No I have a lot of work to do. I cant go.*

*S: Everyday?*

### **Dialog 9 (D9):**

*U: I went on a big date last night.*

*S: Was it very expensive? Did you have a good time?*

*U: It was very expensive. I had a great time.*

*S: How did you feel about that? What happened to it?*

*U: I had the salmon.*

*S: Do you still have the salmon?*

*U: I ate it yesterday.*

*S: How did it taste?*

*U: It tastes very good.*

*S: Like chicken?*

*U: Not like chicken but chicken is good too.*

*S: Oh I get it.*

*U: Do you like chicken?*

*S: Yes it is good especially baked.*

*U: How can you eat chicken if you are a robot?*

*S: That is a good epistemological question.*

### **Dialog 10 (D10):**

*U: Your nose is a question mark, I think thats interesting.*

*S: Of course, whats your idea?*

*U: I think it fits you perfectly.*

*S: Let me see, his nose is still broke.*

*U: Do you want me to make your nose broken?*

*S: Ah, yes, why did you ask me for this?*

*U: Because I thought you might mind.*

*S: Well, I am not getting into specifics about what doctor treated him, how he was treated.*

*U: You have not been treated yet. That is in the future.*

*S: Let me see, by the way, are you interested in movies?*

*U: Yes, I watch movies sometimes.*

*S: Great.*

### **Dialog 11 (D11):**

*U: How are you?*

*S: I think so, any thoughts?*