

Subjective and Objective Evaluation of Conversational Agents

Annika Silvervarg
Linköping University
Linköping, Sweden
annsi@ida.liu.se

Arne Jönsson
Linköping University
Linköping, Sweden
arnjo@ida.liu.se

Abstract

In this paper we present results from an investigation on correlations between subjective and objective evaluation metrics for young people using a conversational agent. The subjective evaluation metrics capture users' experiences of different aspects of conversations with a virtual agent while the objective evaluation metrics are based on an analysis of the actual conversation between the users and the agent. Our study has been conducted using a conversational agent incorporated in a learning environment. The users in the study were pupils in a regular school aged 12 to 14 years.

Our results show that there are no correlations between subjective and objective metrics that are supposed to measure the same aspects, for example, to what extent the system can correctly interpret and give appropriate responses to user utterances. However, users that subjectively like the conversational agent rate its conversational behaviour higher than those that dislikes the system, even though there is no corresponding difference for the objective measures.

Introduction

We are developing a learning environment for young people, most notably teenagers, or almost teenagers. The learning environment includes an embodied agent capable of both task-directed and social interaction with users. The starting point is an existing educational math game (Pareto 2004), in which children train basic arithmetic skills through board games that intertwine game play with learning content through visualisations of arithmetic operations. A crucial part of the game is a pedagogical agent, more specifically a Teachable Agent (TA) (Biswas et al. 2001). The TA is a peer rather than a tutor and the pupil's goal is to teach the agent to play the game. This is mainly done by responding appropriately to different multiple-choice questions posed by the agent during game play, which is called the on-task dialogue. Each question have four candidate answers, one correct, two incorrect, and one "I do not know". These questions are the base for teaching the agent how to play the game.

A novel part of the learning environment is a chat-like written social conversation with the teachable agent, called off-task dialogue. The off-task conversation is a socially oriented chat-like written conversation, where the agent and

student can discuss both domain-oriented topics, such as school and math or off-domain topics like music, friends and family. Reasons for inclusion of such a conversational mode is to increase overall engagement and receptivity of the pupils (Cooper and Baynham 2005), to improve recall of the learning material through emotional engagement (Hamann 2001), to promote trust and rapport-building (Bickmore 2003), and to make pupils feel more at ease with a learning task or topic (Kim et al. 2007). For an extended discussion on the motivation for off-task conversation in a pedagogical setting see Silvervarg, Gulz, and Sjödén (2010). For switching between playing the game and chatting we use the metaphor of regular breaks between lessons in school.

The system uses the metaphor of regular breaks between lessons in school for switching between on-task activities (i.e. playing the game and on-task dialogue) and off-task activities (i.e. social conversation), see Figure 1 for screen shots of the system. Thus, the conversation in our learning environment has a different purpose from those in traditional intelligent tutoring systems, where the conversational agent often acts as a teacher that guides the user through a task, cf. (Graesser et al. 2005; Litman and Forbes-Riley 2006). Our agent has more in common with virtual humans as described by e.g. (Traum et al. 2008), in that it combines social conversation with some task-oriented aspects. As a consequence, the knowledge representation and processing of the dialogue can be less extensive and simpler than in, for instance, traditional task-oriented or tutorial dialogues.

The aim of this paper is to understand the special needs for young people and how to evaluate conversational systems for this very special group. We do this by investigating how objective and subjective evaluations of conversational agents correlate. We first present previous work on objective and subjective measures for evaluation of dialogue system and chatbots, then we describe the off-task conversational abilities of our agent, and finally present and discuss our own empirical findings from a comparison of a subjective evaluation, as collected through questionnaires, and an objective evaluation of tagged logs with actual user conversations with the agent.

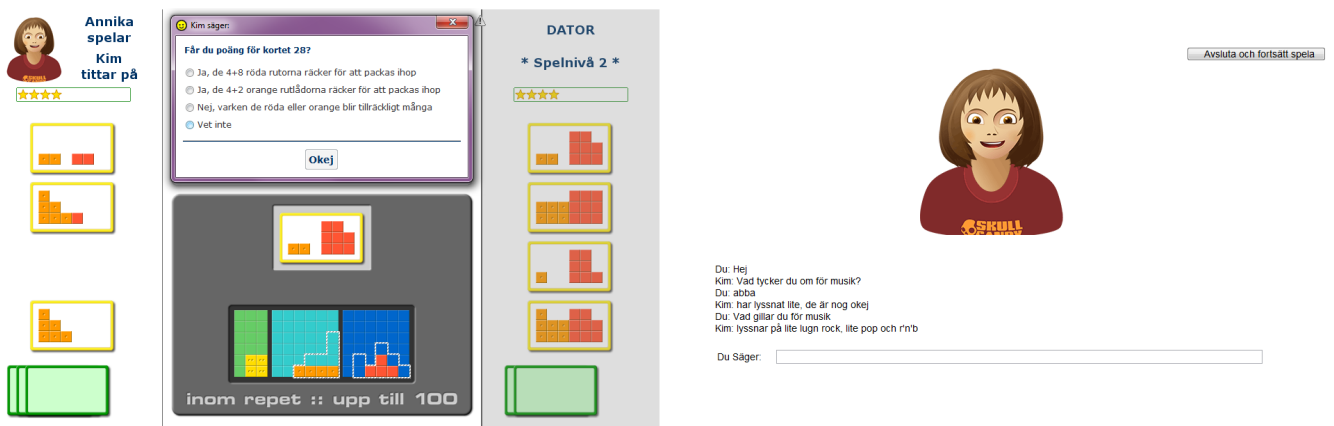


Figure 1: Screenshot of the educational system. On the left side is a screen shot of the educational math game where the agent has asked a multiple choice on-task question. On the right side is a screen shot of the agent engaged in off-task social conversation.

Subjective and objective evaluations of dialogue systems

Evaluation of dialogue systems is mainly done either by distributing a questionnaire to the users trying to reveal their subjective assessment of using the dialogue system or by studying the resulting dialogue. Artstein et al. (2009) call it "soft" numbers versus "hard" numbers and propose a "semi-formal" evaluation method combining the two.

PARADISE (Walker et al. 1998), is one prominent evaluation framework that tries to capture both these perspectives for task-based interactions by combining user satisfaction, task success and dialogue cost into a performance function. PARADISE is developed for task-oriented interactions and requires controlled experiments (Hajdinjak and Mihelič 2006). Walker et al. (1998) claim that it is possible to gather user satisfaction through the performance function. Studies using PARADISE indicate, for instance, that interaction quality is more important than efficiency (Walker, Kamm, and Litman 2000). They also show that there indeed are certain factors that correlate to user satisfaction for task oriented dialogues but that these do not account for all factors correlating to user satisfaction. They show, for instance, that elapsed time is not a good predictor of user satisfaction (Walker, Boland, and Kamm 1999).

For non-task interactions, other factors than task success and dialogue cost are important to achieve user satisfaction, e.g. naturalness. Hung et al. (2009) present a variety of methods that aim at capturing naturalness in interactive conversational agents, or chatbots. Their final method is a framework derived from PARADISE.

Chatbot evaluations have also been conducted using a variety of instruments that tries to combine subjective and objective measures. Shawar and Atwell (2007), for instance, present such investigations, but they never assess the quality of the different measures. Instead they conclude that evaluations should be adapted to user needs and the application at hand.

Foster, Giuliani, and Knoll (2009) performed a study of

subjective and objective evaluation measures using a robot dialogue system for task assembly. They used the objective measures: dialogue efficiency, dialogue quality (interaction problems and number gazes at the objects and robot) and task success. The subjective measures were collected in a questionnaire on opinions of the robot as partner, instruction quality, task success and feelings. Contrary to Walker et al. (1998) they found that dialogue length has a positive contribution to user satisfaction. This is probably because the system is not supposed to find information as efficient as possible, and also because robot-interaction is new and enjoyable. They also found that repetitions lowered user satisfaction. They conclude that user satisfaction factors were affected by other factors than those conveyed in the objective measures.

The Conversational Agent

The conversational dialogue system used in our studies performs off-task conversations with pupils in the context of an educational math game (Silvervarg, Gulz, and Sjöden 2010). The conversational capabilities of the agent have been developed through an iterative process of design - implementation - evaluation, in close cooperation with the target user group, 12 to 14 year old pupils. Each iteration has led to addition of more topics and extended vocabulary, but also more sophisticated strategies for dialogue management (Silvervarg and Jönsson 2010).

In the current version (the fourth iteration) the agent is able to handle greetings, ask and respond to questions and statements on various topics, ask and respond to follow up questions, deal with utterances outside its domain of knowledge, handle flaming/hazing/abuse and other out of scope utterances, know when and how to change topic, and to tell mini narratives, see Figures 2 and 3 for example dialogues.

The off-task conversation is implemented using the open source java-based AIML, Artificial Intelligence Markup

User1: hi
 Kim1: hi
 User2: what are you doing
 Kim2: talking to you
 User3: hahah that's a good one
 Kim3: I do not understand, can you ask something else
 User4: aa
 Kim4: I am so happy today, when I woke up I heard one of my favorite songs on the radio, and it always puts me in a good mood, how are you today?
 User5: do you have a girlfriend?
 Kim5: nope
 User6: ok
 Kim6: how are you?
 User7: it's fine, you
 Kim7: ok
 User8: what do you usually do in the weekends?
 Kim8: can we talk about hobbies instead?
 User9: okey
 Kim9: what do you do during the weekends?
 User10: I practice with the choir ;) you?
 Kim10: aha hang out with my friends
 User11: it's only this weekend that I have practice with the choir
 Kim11: did not understand, what?
 User12: nothing
 Kim12: ok
 User13: do you have an animal
 Kim14: I did not get that, can you ask something else
 User15: so you have a dog?
 Kim15: yep
 User16: what's it's name
 Kim17: Tessy
 User18: okey
 Kim18: What type of music do you like?
 User19: I like pop and r&b you?
 Kim19: mhm
 User20: What music do you like?
 Kim20: I like rnb and easy rock
 User21: okey
 Kim21: I listen to rock and rnb I think that's the sweetest music
 User22: like I di
 Kim22: I think coldplay are so amazingly good, who's your favorite artist?
 User23: I like , b.o.b coldplay ,akon chrisbrown
 Kim23: mhm okey mhm

Figure 2: Example dialogue², where the user has very positive opinions of the conversational agent.

User1: hi there u doin
 Kim1: talking to you
 User2: okey how old are you?
 Kim2: I am 11 year
 User3: want to go steady with me?
 Kim3: I do not understand, ask something else
 User4: are you stupid?
 Kim4: Haha come on, lets talk about something else
 User5: are you gay?
 Kim5: can we talk about hobbies instead?
 User6: do you like the number 6?
 Kim6: did not understand, what?
 User7: can we hang out today?
 Kim7: hm, lets talk about something else
 User8: diss!!!!!!!!!!!!!!!!!!!!
 Kim8: this weekend I am going to the movies. I like to forget about everything and disappear into another world, in case it's a good movie. what about you?
 User9: a together with your mom
 Kim9: okey

Figure 3: Example dialogues where the user has very negative opinions of the conversational agent.

Language³, interpreter called programD⁴. The interpreter supports the use of substitutions, i.e. the opportunity to exchange words or phrases in the user input with other words or phrases. Otherwise only standard AIML has been used.

AIML works on the surface level of utterances, and map user utterances, *<pattern>*, to system responses, *<template>*, which together form a *<category>*. Patterns can consist of words, which in turn consist of letters, numerals, and the wildcards *_* and ***, which function like words. Templates consist in their simplest form of only plain text. However, a template can include AIML tags that allow it to set or get data in variables and predicates, give conditional responses, choose a random response from a set of responses, or recursively call the pattern matcher to insert the responses from other categories. AIML also allows for handling a limited context through the optional tags *<that>*, which refers to the systems last utterance, and *<topic>*, which can span multiple exchanges.

To deal with the variation in user input, synonyms are handled using substitutions, and grammatical variants through several different patterns for the same type of question and topic. The agent's replies are often randomly chosen from a set of 3-5 variants. To be able to correctly respond to follow-up questions and answers to questions posed by the agent, *<that>* and *<topic>* are used. To deal with recurring types of utterances, such as greetings, hazings, and flammings a number of variables are used to keep track of repetitions. To be able to choose new topics the agent has a topic model implemented as a set of AIML predicates including 17 topics that are linked to questions or narratives.

The conversational behaviour is described by a dialogue grammar. The dialogue acts used for the conversation differ

³www.alicebot.org

⁴www.aitools.org

from task-oriented dialogue acts, c.f. Bunt et al. (2010), as our agent is not supposed to carry out a task as efficiently as possible, nor are tutoring-specific dialogue acts, c.f. Litman and Forbes-Riley (2006), applicable as the teachable agent is to be perceived as a pupil rather than a tutor or teacher. The conversational behaviour more resembles that of virtual humans (Traum et al. 2008). Consequently the dialogue acts include task-related as well as dialogue acts used in virtual humans. They comprise: Gr (Greeting), Q (Question), A (Answer), Ack (Acknowledgement), Follow Up (FU), Narrative (N), Not Understood (NU), Not Understood Answer (NUA), Abuse (Ab), Abuse Answer (AbA), and Laughter (Ha). Figure 4 depicts the dialogue grammar based on the dialogue capabilities and dialogue acts described above. Aspects of dialogue behaviour is described in more detail in the following sections.

```
Greet ::= GrU GrA [GrU AgentQ|AgentN]
AgentN ::= NA [AckU AgentQ]
AgentQ ::= QA AU [AgentAck]
AgentQ ::= QA AU [AckA UserFU]
AgentQ ::= QA AU FUA [UserAck]
AgentQ ::= QA UserAFU
UserAFU ::= AU FUA AA [UserAck]
UserFU ::= FUU AA [UserAck]
UserQ ::= QU AA [UserAck]
UserQ ::= QU AgentAFU
AgentAFU ::= AA FUA AU [AgentAck]
UserAck ::= AckU AgentAck
AgentAck ::= AckA [AckU AgentN|AgentQ]
Abuse ::= AbU AbAA1 [Abuse2]
Abuse2 ::= AbU AbAA2 [Abuse3]
Abuse3 ::= AbU AgentN|AgentQ [Abuse4]
Abuse4 ::= AbU AbAA4
NotUnderstand ::= NUU NUAA1 [NotUnderstand2]
NotUnderstand2 ::= NUU NUAA2 [NotUnderstand3]
NotUnderstand3 ::= NUU |AckU AgentN|AgentQ
Laughter ::= LU LA [LU AgentN|AgentQ]
```

Figure 4: Dialogue Grammar. The dialogue grammar describe the agents conversational behaviour. The subscript letter is used to indicate if the user (U) or the agent (A) performs the dialogue act and the superscripts (1, 2, 3, 4) on the agent responses to Abuse and NotUnderstand utterances indicate that the response comes in different variants. Brackets are used for optional items and alternatives are separated by |.

Initiative

The agent is designed to allow for a mixed-initiative dialogue where both agent and user can initiate new topics and ask general follow up questions, e.g. "what about you". A new topic is initiated in situations where interaction is repeated or the user does not take the initiative. For example, when greetings, acknowledgments or laughs are repeated the agent asks a question or tells a mini narrative, as exemplified in Figure 5⁵.

⁵All dialogues are translated from Swedish by the authors. Dialogue acts are shown in parenthesis after each utterance.

```
User: Hi (GrU)
Kim: Hi (GrA)
User: Hello (GrU)
Kim: What type of music do you like? (QA)
User: much u? (AFUU)
Kim: everything but dance band music hehe (AA)
User: hehe (UL)
Kim: lol (AL)
User: lol (UL)
Kim: I listen to rock and rnb I think that's the sweetest music (AN)
```

Figure 5: Example of Kim taking the initiative.

The agent can also take the initiative when the user repeatedly abuse Kim, or there have been many consecutive user utterances that have not been understood, as described in the two following sections.

Repair strategy

A recovery strategy to deal with the utterances that the system is not able to interpret correctly is needed to help the user understand what they can talk about and avoid losing track of the conversation. The strategy used is similar to the one used by SGT Star (Artstein et al. 2009); first a simple *clarification* then, a *stall*, followed by a *move on*, and finally a *prompt* that directs the user to a new topic. In our system the agent stall is a general prompt for the user to initiate a new topic, if this fails the agent proposes a topic, and as last resort takes the initiative and asks a question or tells a narrative, Figure 6.

```
User: fgd (NUU)
Kim: I did not understand that (NUAA1)
User: cb (NUU)
Kim: I did not catch that, can we talk about something else? (NUU2)
User: s (NUU)
Kim: Lets talk about school (NUAA3)
User: s (NUU)
Kim: My school is okey, a little to much homework, wbu? (NUU4)
```

Figure 6: Example of how Kim handles utterances that are not understood

Hazing/Flaming/Testing strategy

In the dialogues in the corpus, hazing, testing and flaming occurs rather frequently. To fend this off, the agent recognizes such utterances and use a strategy similar to the repair strategy where the two first responses are general, followed by an attempt to change the subject, see Figure 7.

Subjective and objective evaluations

The conversational agent was objectively and subjectively evaluated at a Swedish School, where pupils from three classes, 12-14 years old, used the learning environment with the conversational agent during three lectures. Data from questionnaires and dialogue logs from a total of 19 pupils, 8

User: you are ugly (Ab_U)
 Kim: eh, lets talk about something more interesting (AbA_A^1)
 User: you smell (Ab_U)
 Kim: come on, lets talk about something fun (Ab_U^2)
 User: ugly (Ab_U)
 Kim: Lets talk about music (Ab_A^3)
 User: ugly (NU_U)
 Kim: Stop that, or I will tell a teacher (Ab_U^4)

Figure 7: Example of how Kim handles abuse

females and 11 males, were used in this study. The pupils played the game for about a total of 120 minutes, and after every second game session a break was offered. During the first three breaks the pupils had to chat with the agent until the break ended, after that chatting was optional.

Questionnaire

After the final session a questionnaire was distributed to the pupils. The questionnaire is partly based on SASSI (Subjective Assessment of Speech System Interfaces) (Hone and Graham 2000) and CCQ (The Communication Competence Questionnaire) (Monge et al. 1982). It consists of Likert items scaled from 1 (Strongly disagree) to 7 (Strongly agree), see Figure 8. The questionnaire items were chosen to capture aspects of the agents conversational abilities, e.g. that the agent understood user utterances and could give appropriate responses, as well as the users experience of conversing with the agent, e.g. naturalness and likeability.

Dialogue Coding Scheme

The coding scheme used to analyse the dialogue logs is based on the coding schemes used by Robinson, Roque, and Traum (2010) with a set of codes characterizing the users dialogue action and another set of codes that evaluates the agent’s response. For the investigations presented in this paper we only use a subset of the codes in the top layer of Robinson, Roque, and Traum (2010), Table 1.

Table 1: Dialogue action codes

Code	Description
D	General dialogue functions, e.g. Greeting, Closing, Politeness
H	Hazing, Testing, Flaming
Q	Information Request
R	Requests, in our case for help and clarification
A	Answer to agent utterances
S	Statements
G	Incomprehensible, typically nonsense or empty utterances

We also use a category for agent initiatives, I, and one for repeated agent utterances, REP.

Table 2: Agent response values

Code	Value
3	Correct interpretation
2	Partially appropriate
RR	Request Repair
1	Inappropriate response

Speaker, Utterance	Code
User are you free on tuesday	Q-RR
Kim I did not understand what you said, can we talk about something else	
User hi	D-3
Kim Hi	
User are you good at school	Q-3
Kim rather good	
User need	H-3
Kim maybe we can talk about hobbies instead	I
User I like football	S-RR
Kim what, how do you mean?	
User I like kicking the balls	S-RR
Kim hmm, I have difficulties understanding what you say, can we talk about something else	
User do you have a problem with Swedish	Q-2
Kim this weekend I am going to the movies. It is great to forget everything and just disappear into another world, at least if it is a good film, how about you?	I
User may I join	Q-2
Kim What are you doing at the weekends?	I

Figure 9: Example dialogue, translated from Swedish by the authors.

The agent’s responses are evaluated using a modified version⁶ of the coding scheme used by Robinson, Roque, and Traum (2010), Table 2. Correct interpretation means that the system understood the user and responded correctly. Partially appropriate, 2, means that the user’s utterance is uninterpretable but the system’s response is to initiate a new topic, see the dialogue in Figure 9. Request Repair is when the system does not understand and asks for a clarification or that the user changes topic. Inappropriate response, finally, is when the system responds erroneously.

Metrics

As the purpose of this research is to investigate how subjective and objective measures correlate, we need to map the subjective and objective measures used in our study. Table 3 depicts our mappings.

Some mappings are straightforward, such as Questionnaire item 2, Q2 *Kim understood what I said*, see Figure 8, which is correlated with the proportion of correctly interpreted responses from the agent. The amount of nonsense, $n(G)$, is removed from the total, N , as such utterances never

⁶In our corpus we do not have any instances of “No response” so we end up with four values for the agent’s response.

Questionnaire item	N	Min.	Max.	M	SD
1. Kim’s answers often surprised me	19	5	3	4.05	2.27
2. Kim understood what I said	19	5	2	3.37	2.01
3. I could fix misunderstandings if I wanted to	19	4	7	4.79	2.39
4. Kim was a good listener	19	6	5	4.05	2.48
5. I would like to talk to Kim again	19	3	5	4.32	2.29
6. Kim expresses her ideas very clearly	19	4	5	4.47	2.27
7. Kim mostly says the right thing at the right time	19	4	4	4.05	2.32
8. Kim is easy to talk to	19	3	5	4.37	2.03
9. I liked to talk to Kim	19	3	5	4.37	2.22
10. I could control the interaction with Kim	19	3	5	4.42	2.17
11. It was easy to understand how to talk so that Kim should understand	19	3	3	4.00	2.13
12. It felt natural to talk to Kim	19	5	2	3.79	2.25
13. Sometimes I lost track of the conversation	19	2	4	4.37	1.92
14. It was frustrating to talk to Kim	17	1	7	5.12	2.06
15. It was hard to know what to talk about with Kim	19	3	4	4.12	2.23
16. Kim often repeated herself	19	12	1	2.05	1.78
17. Sometimes I wondered if I used the right word	19	1	7	4.37	2.22
18. I always knew what I could say to Kim	18	5	4	4.17	2.38

Figure 8: Questionnaire items and statistics from the evaluation. *Max.* denoted the number of highest score, *Min.* the number of lowest score, average (M) and standard deviation (SD). The scores on the negatively loaded items (13-17) were transformed so that a high score is positive for the dialogue system and a low score is negative for the system.

Table 3: Mapping of subjective and objective measures. N is the total number of utterances and $n(x)$ denotes the number of utterances tagged as category x , $|$ denotes or and $X - Y$ denotes a turn-taking, e.g. $n(Q - 3)$ denotes the number of user questions, Q , followed by a correct agent response, 3.

Description	Questionnaire	Dialogue rating
Correct interpretation	Q2	$\frac{n(3)}{N - n(G)}$
Appropriate response	Q7	$\frac{n(2) + n(3) + n(RR)}{N}$
Repetition	Q16	$\frac{n(REP)}{N}$
Control	Q10	$\frac{n(I)}{N}$
Coherence	Q13	$\frac{n(Q S-1 RR)}{N}$
Habitability	Q11, Q18	$\frac{n(D Q S-3) + n(D Q S-2)}{n(D Q S)}$

can be interpreted by the system and always result in 2 or *RR*.

Slightly more general is item Q7 *Kim mostly says the right thing at the right time* which should correlate with the proportion of answers that are appropriate, i.e. including also partially appropriate responses, 2, and requests for repair, *RR*.

The proportion of repetitions, *REP*, directly corresponds to item Q16 *Kim often repeated herself*.

To correlate the user’s sense of control, captured in item Q10 *I could control the interaction with Kim*, we use the proportion of initiatives the system takes, *I*-tags, i.e. if the system takes initiative we consider that as lack of user control. Normally the user has control, the system mainly takes the initiative when the user do not seem to want to control the interaction, as described in section .

The questionnaire item Q13 *Sometimes I lost track of the conversation* is correlated with the proportion of questions or statements that the system has misinterpreted and given

faulty answers to, or utterances where the system responds that it has not understood. Such responses do not contribute to the flow of the conversation and is assumed to interrupt the users’ track of conversation.

One important property of our system is habitability which is captured through the items Q11 *It was easy to understand how to talk so that Kim should understand* and Q18 *I always knew what I could say to Kim*. There is no obvious utterance type that directly correlates to habitability. We believe, however, that habitability can be correlated with the proportion of correct responses from the system to the users’ questions, *Q*, statements, *S*, and greetings, closings and politeness, *D*, since this indicates that the user has been able to express such utterances in a way that the system can understand. Correct response does not necessarily mean that the system’s interpretation is correct, a correct chat conversation also includes appropriate responses (tagged 2), see Figure 9. Such sequences depict conversations that flow naturally and as the user often has the initia-

Table 4: Subjective measures with mean (M), standard deviations (SD), and number of extreme values Min. or Max.. M and t are also shown for the two subgroups of pupils that like (L) and do not like (NL) the conversational agent.

Questionnaire item	N	Min.	Max.	M	SD	M _L	M _{NL}	t
Likeability (Q5, Q9)	19	2	5	4.34	2.23	5.83	1.79	<0.001
Naturalness (Q8, Q12)	19	3	1	4.08	2.04	5.25	2.07	<0.001
Correct interpretation (Q2)	19	5	2	3.37	2.01	4.33	1.71	<0.001
Appropriate Response (Q7)	19	4	4	4.05	2.32	5.25	2.00	<0.001
Repetition (Q16)	19	1	12	2.05	1.78	1.58	2.86	<0.1
Control (Q10)	19	3	5	4.42	2.17	5.75	2.14	<0.001
Coherence (Q13)	19	2	4	4.37	1.92	3.83	5.29	<0.1
Habitability (Q11, Q18)	19	3	2	4.03	2.18	5.33	1.79	<0.001

Table 5: Objective measures with mean (M) and standard deviations (SD). M and t are also shown for the two groups like (L) and do not like (NL) the conversational agent.

Dialogue coding	N	Min	Max	M	SD	M _L	M _{NL}	t
Correct interpretation	19	0.32	0.76	0.58	0.12	0.57	0.59	-
Correct Response	19	0.88	1	0.96	0.03	0.93	0.96	-
Repetition	19	0	0.16	0.09	0.04	0.11	0.06	<0.05
Control	19	0.12	0.38	0.29	0.06	2.15	1.86	<0.1
Coherence	19	0.09	0.39	0.21	0.09	0.22	0.20	-
Habitability	19	0.16	0.75	0.49	0.15	0.50	0.48	-
Flaming/Hazing	19	0	0.55	0.11	0.14	0.07	0.18	<0.05

tive we believe that it is an indication of habitability. The reason for not dividing by the total number of utterances, N , is that N includes all Hazing/Flaming/Testing H and Non-interpretable G utterances, which varies between users, and these are not relevant since the user have not seriously tried to communicate with the agent in those turns of the dialogue.

Results

The results from our two evaluations will be presented briefly before presenting the correlations.

Subjective evaluation

Table 4 shows the results from the subjective evaluation, where items from the questionnaire has been reduced to a number of factors that capture various aspects of how the agent’s conversational abilities and the dialogue with the agent is experienced. As can be seen the overall impression of the conversational agent is that it is neither very good or bad as many measures have values around 4, for example likeability ($M = 4.34$) and naturalness ($M = 4.08$). The agents conversational abilities are also neither good nor bad (correct interpretation $M = 3.37$, appropriate response $M = 4.05$), and it is neither hard nor easy to know how to interact with the agent (habitability $M = 4.03$).

However, there is a fairly large variation as indicated by standard deviations around 2 and in many cases high frequencies of both 1s and 7s. Therefore we decided to further investigate subgroups of users. We divided our users in two groups based on how well they liked to chat with the agent, based on the likeability ratings of questionnaire items

5 I would like to talk to Kim again and 9 I liked to talk to Kim. As seen in the right hand sided columns in Table 4, there are significant differences between the group that like to chat (L) and those who do not like to chat (NL) for all factors except repetition, concerning how they perceive the conversation with the agent.

Objective measures

Table 5 shows the corresponding values for the metrics used to evaluate the dialogues. Since these are calculated as fractions, all values range from 0 to 1. In theory all metrics can take values from 0 to 1 but in practice the dialogue is designed so that the agent never can take the initiative in all turns, thus *Control* has a maximum value closer to 0.5. While there are large variation between the max and min values for the objective measures, the objective measure differ from the subjective in that the standard deviations are much smaller. For some measures the mean falls in the middle, e.g. correct interpretation and habitability, but others are more on the extreme end of the scale, e.g. appropriate response. Looking at the subgroups based on whether they liked the chat or not, the only significant difference is that there is more flaming/hazing in NL. From our objective measures it seems that the system handles most utterances appropriately, although many of these are examples of RR, see Table 6.

Correlations between subjective and objective measures

No significant correlations between subjective and objective measures could be found. The correlation coefficients were

Table 6: Proportion of different responses.

Code	Proportion (%)
3	51
2	15
RR	30
1	4

approximately 0.2 for all aspects, except for *Control*, which was 0.3. Looking at the subgroups Like and Not Like only revealed a single correlation between the subjective and objective measures for *Control*, which was 0.7.

The lack of correlations can probably be explained by the fact that although there are large individual differences in the subjective evaluation, between those that like the system and those that do not, see the right most columns in Table 4, there is no corresponding variance in the actual dialogues, i.e. the objective measures are more homogenous, see the right most columns in Table 5.

The number of subjects used in this study is admittedly small and further studies may provide significant correlations between subjective and objective measures. For instance, the questionnaire was distributed after a rather long period (120 mins). In a recent study we limited that to 30 minutes and we are currently investigating how this influences the results from the subjective measures and in turn, the correlations.

Discussion

We have presented results from an investigation on young peoples subjective evaluation of conversational agents and their actual conversation with the system.

Our results show that users that like to use the system experience the agent's conversational abilities as better than users that do not like the system, regardless of how well the system manages the conversation with the user. It, thus, seems that the attitude towards the system is more important than its actual conversational capabilities.

To further investigate this, we have compared the results from a subjective evaluation of a previous version of the system where we used the same items in the questionnaire as in this study. We divide the users into those that liked to use the system and those that did not. We found that for most of the items used in this study, see Table 4, we have a significantly higher rating on the current system for those that like the system, but no such differences for those that do not. Thus, those that like the system also appreciate improvements made to the systems conversational capabilities, whereas those that dislike the system don't care.

To conclude, measures from our objective and subjective evaluation of a conversational agent do not correlate. An implication of this is that data from subjective evaluations cannot be the only source of information to assess conversational agents (maybe if those that dislike the system are removed from the data), but neither can objective measures. However, objective measures are more homogenous and therefore probably better reflect a conversational

system's capabilities. But as they do not correlate with the subjective measures they cannot be used to predict user satisfaction.

References

- Artstein, R.; Gandhe, S.; Gerten, J.; Leuski, A.; and Traum, D. 2009. Semi-formal evaluation of conversational characters. *Languages: From Formal to Natural* 22–35.
- Bickmore, T. 2003. *Relational Agents: Effecting Change through Human-Computer Relationships*. Ph.D. Dissertation, Media Arts & Sciences, Massachusetts Institute of Technology.
- Biswas, G.; Katzlberger, T.; Brandford, J.; D., S.; and TAG-V. 2001. Extending intelligent learning environments with teachable agents to enhance learning. In Moore, J.; Redfield, C.; and Johnson, W., eds., *Artificial Intelligence in Education*. Amsterdam: IOS Press. 389–397.
- Bunt, H.; Alexandersson, J.; Carletta, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Lee, K.; Petukhova, V.; Popescu-Belis, A.; Romary, L.; Soria, C.; and Traum, D. R. 2010. Towards an ISO standard for dialogue act annotation. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.
- Cooper, B., and Baynham, M. 2005. Rites of passage: embedding meaningful language, literacy and numeracy skills in skilled trades courses through significant and transforming relationships. Technical report, National Research and Development Centre for Adult Literacy and Numeracy.
- Foster, M. E.; Giuliani, M.; and Knoll, A. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, 879–887. The Association for Computer Linguistics.
- Graesser, A.; Chipman, P.; Haynes, B.; and Olney, A. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education* 48:612–618.
- Hajdinjak, M., and Mihelič, F. 2006. The PARADISE evaluation framework: Issues and findings. *Computational Linguistics* 32(2):263–272.
- Hamann, S. 2001. Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences* 5(9):394–400.
- Hone, K., and Graham, R. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6(3/4):287–305.
- Hung, V.; Elvir, M.; Gonzalez, A.; and DeMara, R. 2009. Towards a method for evaluating naturalness in conversational dialog systems. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA*, 1236–1241.

- Kim, Y.; Wei, Q.; Xu, B.; Ko, Y.; and Ilieva, V. 2007. athgirls: Increasing girls' positive attitudes and self-efficacy through pedagogical agents. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*.
- Litman, D., and Forbes-Riley, K. 2006. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*.
- Monge, P. R.; Bachman, S. G.; Dillard, J. P.; and Eisenberg, E. M. 1982. Communicator competence in the workplace: Model testing and scale development. In *Communication Yearbook*, 5. Beverly Hills, CA:Sage. 505–528.
- Pareto, L. 2004. The squares family: A game and story based microworld for understanding arithmetic concepts designed to attract girls. In *In World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 2004, 1567–1574.
- Robinson, S.; Roque, A.; and Traum, D. R. 2010. Dialogues in context: An objective user-oriented evaluation approach for virtual human dialogue. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.
- Shawar, B. A. A., and Atwell, E. S. 2007. Chatbots: are they really useful? *LDV-Forum* 22:31–50.
- Silverbarg, A., and Jönsson, A. 2010. Towards a conversational pedagogical agent capable of affecting attitudes and self-efficacy. In *Proceedings of the Second Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity, Bucharest, Romania*.
- Silverbarg, A.; Gulz, A.; and Sjöden, B. 2010. Design for off-task interaction – rethinking pedagogy in technology enhanced learning. In *Proceedings of the 10th IEEE Int. Conf. on Advanced Learning Technologies, Tunisia*.
- Traum, D. R.; Swartout, W.; Gratch, J.; and Marsella, S. 2008. A virtual human dialogue model for non-team interaction. In Dybkjaer, L., and Minker, W., eds., *Recent Trends in Discourse and Dialogue*. NY: Springer. 45–67.
- Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1998. Paradise: A framework for evaluating spoken dialogue agents. In Wahlster, M. M. . W., ed., *Readings in Intelligent User Interfaces*. Morgan Kaufmann.
- Walker, M.; Boland, J.; and Kamm, C. 1999. The utility of elapsed time as a usability metric for spoken dialogue systems. In *Proceedings of ASRU*, 317–320. Citeseer.
- Walker, M.; Kamm, C.; and Litman, D. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering* 6(3):363–377.