



NECA

NET ENVIRONMENT FOR EMBODIED EMOTIONAL
CONVERSATIONAL AGENTS

D9e: Report on Demonstrator Evaluation Results

ErichGstrein, Christoph Schmotzer, Sysis

Brigitte Krenn, OFAI



Version: 0.1

Date: 28. July 2004

Project ref. no.	<i>IST-2000-28580</i>
Project title	NECA: A Net Environment for Embodied Emotional Conversational Agents
Deliverable status	Public
Contractual date of delivery	May 2004
Actual date of delivery	July 2004
Deliverable number	D9e
Deliverable title	Report on Demonstrator Evaluation Results
Type	Report
Status & version	Final
Number of pages	18
WP contributing to the deliverable	WP 9
Task responsible	Sysis/OFAI
Author(s)	Erich Gstrein, Christoph Schmotzer, Brigitte Krenn,
EC Project Officer	Erwin Valentini
Keywords	demonstrator evaluation
Abstract dissemination) (for	Evaluation of the beta version of the NECA demonstrators Socialite and eShowroom in a field trial.

Erich Gstrein, Christoph Schmotzer

Sysis interactive simulations AG

Hasnerstraße 123

A-1160 Wien

Austria

Brigitte Krenn

Austrian Research Institute for Artificial Intelligence

Freyung 6/6

1010 Wien

Austria

1. Introduction

While the initial versions of the NECA demonstrators Socialite and eShowroom were accessible only for a restricted audience, the β versions have been made available to the unrestricted public. The β version of the Socialite demonstrator was linked to the largest Austrian student portal 'cycamp.at' (www.cycamp.at) in March 2004 and was set online on the 19.3.2004. (Note, Socialite is integrated in the community building application 'derSpittelberg' which creates a virtual student community socio-culturally anchored in the Viennese Spittelberg area. 'derSpittelberg' is an application of the Sysis NetLife community building tool.) As basis for the analysis presented in this document, a snapshot of the 'derSpittelberg' database was taken on the 21.7.2004, covering 124 days online. The β eShowroom was set online in January 2004 (www.eshowroom.org), and a snapshot of the database was taken covering the time period from 1.2.2004 until 25.5.2004, i.e., 140 days online. It must be noted that the applications have not been particularly advertised. There is only a link from 'derSpittelberg' at cycamp to the eShowroom. Thus we have (deliberately) generated a situation for a field test under aggravating circumstances as regards the number of users attracted. In other words, we have set a baseline in terms of a minimum of user attraction, in order to fathom the minimum potential of applications featuring animated conversation among embodied conversational characters.

Two parallel questionnaires were designed for the demonstrators, asking whether the animated presentation was fun, whether body motion and facial expression did match, the voices of the animated characters were appropriate, etc. Moreover, for direct comparison of two applications the same semantic differential was included in the two questionnaires assessing the overall impression of the user on the animated dialogue. The questionnaire was presented to the user after a completed dialog session. In the case of Socialite this means after each dialogue scene presented to the user the questionnaire pops up. In the case of eShowroom where the user can change the settings for the presentation the user is presented with the questionnaire after the last presentation selected.

2. Questionnaires

In the following, we first present the eShowroom questionnaire and then the Socialite questionnaire as they appear to the user. (Note, eShowroom is an English application whereas Socialite is a German application.) As already stated the questionnaires were built in parallel only diverging in the language and some application specific formulations. The questionnaires were designed in collaboration of DFKI, OFAI, Sysis and the Austrian usability lab CURE (www.cure.at).

Each questionnaire comprises eight question areas assessing different aspects of the applications, which are listed in the following:

- 1) how the user enjoyed the application
- 2) how the user perceived the match between facial expression, body movements and speech
- 3) how the users perceived the voices of the characters
- 4) the user's overall (positive/negative) assessment of the animated dialogues

5) and 6) assess the match between the agent behaviour and the user expectations on the agent behaviour. In the case of eShowroom it was assessed whether the two car sales presentation characters, Tina and Ritchie, performed according to the parameters set by the user. (Note, the user in eShowroom can define who takes on the role as seller and who is the buyer, how the mood of the individual characters is and whether a character is polite or impolite. These settings and the selected value dimensions of the car such as comfort, safety etc. drive the generation of the animated presentation scene played to the user.) In the case of Socialite it was asked whether the behaviour of the user avatar has matched the user expectations (5), and whether the other avatar was experienced as friendly or likeable (6).

7) is a collection of (optional) questions on gender, age, proficiency with computers, experience with animated characters.

8) allows for free text input

NECA eShowroom Questionnaire

Please consider the following statements carefully and rate your agreement with each one. The completion of this questionnaire will take you no longer than three minutes but your feedback will help us to improve eShowroom.

When finished, please press the **Submit** button at the bottom of this page. The information you provide is kept completely confidential.

I enjoyed watching the car sales dialogues. **Agree** **Disagree**

The body movements and the facial expressions of the virtual actors matched their spoken utterances. **Agree** **Disagree**

The voices of the virtual actors sounded natural and appropriate. **Agree** **Disagree**

Please decide between the following pairs of adjectives according to your subjective impression of the whole animated dialogue: The animated dialogue was?

Outstanding	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Second-rate
Exclusive	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Standard
Impressive	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Nondescript
Unique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Ordinary
Innovative	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Conservative
Exciting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Dull
Interesting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Boring

Tina performed according to the parameters that I had set before the **Agree** **Disagree**

presentation started.

Ritchie performed according to the parameters that I had set before the presentation started.

Agree Disagree

Please provide us with the following voluntary information.

gender:

age:

What is your experience with computers?

What is your experience with animated characters?

If you have further comments or if you think that something important was not covered by these questions please give us your feedback.

powered by



NECA Spittelberg Fragebogen

Wir - das Spittelberg Team - möchten Dich bitten, folgende Aussagen sorgfältig durchzulesen und deine Bewertung dazu abzugeben. Du benötigst für die Beantwortung dieses Fragebogens lediglich drei Minuten. Damit hilfst Du uns bei der Verbesserung von 'Spittelberg'.

Nach der Abgabe deiner Bewertungen drück' bitte den **Speichern** Knopf am Ende dieser Seite. Es versteht sich von selbst, daß Deine Daten streng vertraulich behandelt werden.

Das Ansehen der Dialoge hat mir Spaß gemacht.

Stimme zu Stimme nicht zu

Die Bewegungen und die Gesichtsausdrücke der virtuellen Personen stimmten mit den gesprochenen Äußerung überein.

Stimme zu Stimme nicht zu

Die Stimmen der virtuellen Personen klangen passend und natürlich.

Stimme zu Stimme nicht zu

Bewerte bitte den gesamten animierten Dialog anhand der angeführten Eigenschaftspaaren: Der animierte Dialog war ...

Hervorragend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Zweitklassig
Exklusive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Standard
Eindrucksvoll	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unbestimmbar
Einzigartig	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gewöhnlich
Innovativ	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Konservativ
Aufregend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Platt
Interessant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Langweilig

Das Verhalten meines Avatars hat meinen Vorstellungen entsprochen.

Stimme zu Stimme nicht zu

Ich fand den Gesprächspartner meines Avatars sympathisch.

Stimme zu Stimme nicht zu

Weiters würden wir noch gerne folgendes von Dir wissen.

Geschlecht:

Alter:

Wie gut kennst Du dich mit Computern aus?

- bitte auswählen -

Welche Erfahrungen hast Du mit animierten Charakteren?

- bitte auswählen -

Falls Du weitere Anmerkungen machen willst oder Du der Meinung bist, daß wichtige Aspekte von diesen Fragen nicht abgedeckt wurden, so laß uns dies wissen.

A rectangular text area with a light gray background and a thin black border. It contains no text. On the right side, there are two vertical scroll arrows (up and down). On the bottom side, there are two horizontal scroll arrows (left and right).

Speichern

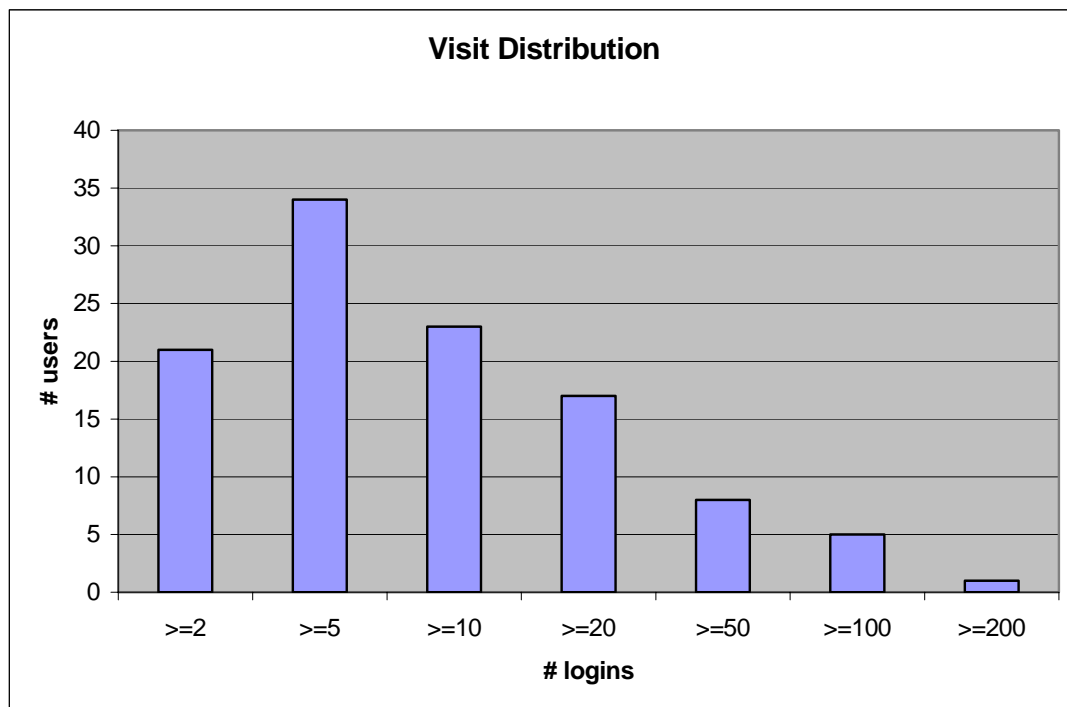
3. Basic Statistics: 'derSpittelberg'

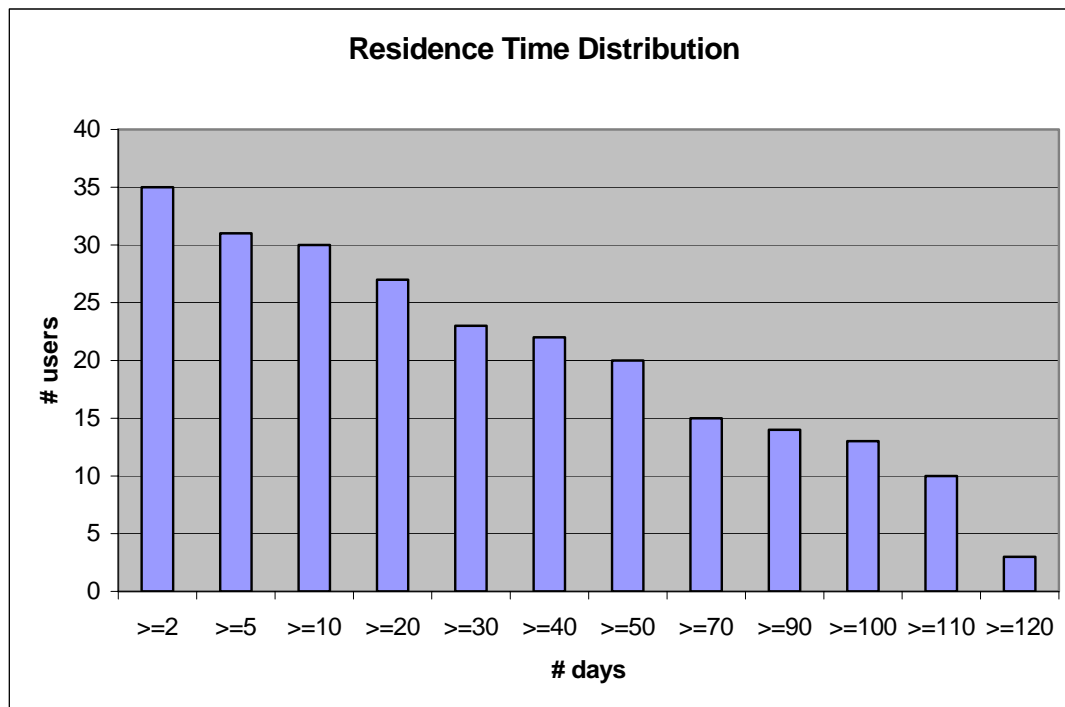
In 'derSpittelberg', being a Sysis NetLife application, a number of usage data is logged. Some of these are used for a general view on the user group of β Socialite which is presented in the following.

First of all we give some basic information:

- The evaluation/online period of the Socialite demonstrator was 124 days
- 66 users registered in 'derSpittelberg' community
- 1488 logins were made, this is ~22 logins per user over the whole period of 124 days. The maximum was of an individual user was 241 visits over a period of 122 days.
- On average 12 logins were made per day. The maximum was 75, the minimum was 0 login on 6 different days.
- The average residence time of the users in 'derSpittelberg' was 34 days – (calculated as the average value of 'last visit – first visit').

The following tables show the distributions of the number of users over the performed logins and the residence time.





As shown in the table on Visit Distribution, the number of logins per user peaks between 5 to 10 logins, i.e., approximately half of the users logged in (visited their avatar) at least 5 times. A small fraction of users shows much higher logins, see for instance the group with at least 100 logins per user. The table on Residence Time Distribution clearly shows that with increasing residence time (number of days in the community) the number of users decreases.

Summing up, one of the most prominent inferences that can be drawn from the data presented above is: although the number of registered users is not particularly high, the users like to take part in the application and frequently return to the community.

For eShowroom we have no comparable data of this kind, because it is a completely different type of application where a single user is more likely to go there once, try out the different settings concerning the value dimensions of the goods (car) presented and play with the personality and mood settings of the characters.

4. Questionnaire Results

In the following, we discuss the results from the questionnaires.

4.1 Methods

As it is typical for unforced non-laboratory questionnaire studies, the return rate for the eShowroom and the Socialite questionnaires was very low. Because of the small number of collected questionnaires – 11 from eShowroom and 17 from the Socialite – statistical analysis was out of question. Therefore we decided to look at the distributions and mean values of the answer values. Thus the following evaluations need to be taken with caution, as they can only give a flavour or ideas on tendencies which need to be verified in further experimentation (qualitative or quantitative). The insights gained from such small user groups, however, are valuable information in a user centered design approach.

4.2 Socialite: Some General Data

- 17 questionnaires were collected, written from 8 distinct users. We use 17 as basis. This is justified by the fact that the filled in questionnaires relate to different animated dialogue scenes. In other words, there are no fully identical dialogue scenes generated in the NECA system. This is due to a random component in the multimodal natural language generator.
- It is remarkable that all questionnaires (per user) were produced within a single session. Later visits did not lead to any new records of the same user! This means, that only 0,5% of the logins led to a filled in questionnaire.
- It is also remarkable that the users who rated the dialogues (produced a filled in questionnaire) had a higher performance in terms of logins and persistence than the average user. In other words, only those users highly involved with the application were willing to fill in a questionnaire or vice versa.
- Most users (88% of those who filled in the questionnaires) stated to use animated characters on a regular basis, whereas 6% rated themselves as an expert and another 6% told, that this was the first time meeting an animated avatar.
- Most users (88% of those who filled in the questionnaires) stated to have ‘expert knowledge’ about computers, and 12% had ‘some knowledge’.
- 70% of the users were males.
- 58% were between 20 and 29 years and 42% were between 30 and 39 years old.

4.3 Socialite: Evaluation of the Questions

- Question 1 (Overall enjoy): 47% enjoyed the dialouges, 29% had a neutral opinion and 24% gave a negative rating. Males gave better ratings than females – no male gave a ‘worst case’ rating (-2).

enjoy	-2	-1	0	1	2
%M	0	25	25	16,6666667	33,3333333
%F	20	0	40	20	20
%ALL	5,88235294	17,6470588	29,4117647	17,6470588	29,4117647

Distribution of answers (percent values)

all	male	female	age: 20-29	age: 30 -39
0,47	0,58	0,2	0	1,14

Mean values of answers to Question 1

- Question 2 (Movement): about 48% gave a positive, 40% a neutral and only 12% a negative rating. Males gave slightly better ratings; especially by applying the highest rates.

movements	-2	-1	0	1	2
%M	0	8,33333333	41,6666667	16,6666667	33,3333333
%F	0	20	40	40	0
%ALL	0	11,7647059	41,1764706	23,5294118	23,5294118

Distribution of answers (percent values)

all	male	female	age: 20-29	age: 30 -39
0,59	0,75	0,2	0,2	1,14

Mean values of answers to Question 2

- Question 3 (Voice): 82% of the users gave negative ratings, 12% were neutral and only 5% were positive. No male made positive ratings – but 20% of the females gave a '+'. But males and females gave an accordantly negative rating.

voice	-2	-1	0	1	2
%M	8,33333333	75	16,6666667	0	0
%F	20	60	0	20	0
%ALL	11,7647059	70,5882353	11,7647059	5,88235294	0

Distribution of answers (percent values)

all	male	female	age: 20-29	age: 30 -39
-0,88	-0,92	-0,8	-1	-0,7

Mean values of answers to Question 3

- Question 4:

	all	male	female	20-29	30 -39
Hervorragend - Zweitklassig	0,24	0,42	-0,20	-0,5	1,28
Exklusive - Standard	0,29	0,5	-0,20	-0,6	1,15
Eindrucksvoll-Unbestimmbar	0,18	0,42	-0,4	-0,8	1,15
Einzigartig - Gewöhnlich	0,47	0,75	-0,40	-0,5	1,8
Innovativ - Konservativ	0,59	0,67	0,40	0	1,14
Aufregend - Platt	0,65	0,83	0,2	-0,7	2
Interessant - Langweilig	0,65	0,83	0,2	-0,2	1,8

Mean values of answers to Question 4 (semantic differential)

- Question 5 (My Avatar): The answers a very uniformly distributed, but all in all a slightly negative impression is described. 47% gave a negative, 18% a neutral and 35% a positive rating.

myAvatar	-2	-1	0	1	2
%M	16,6666667	33,3333333	16,6666667	8,3333333	25
%F	20	20	20	40	0
%ALL	17,6470588	29,4117647	17,6470588	17,6470588	17,6470588

Distribution of answers (percent values)

all	male	female	age: 20-29	age: 30 -39
-0,12	-0,08	-0,2	-0,3	0,14

Mean values of answers to Question 5

- Question 6 (Other Avatar symp.): 47% gave a negative , 6% a neutral and 47% a positive rating on the sympathy of the other avatar. Interestingly, females gave very negative ratings (60% neg. and 40% worst case!).

otherAvatar	-2	-1	0	1	2
%M	33,3333333	8,3333333	8,3333333	25	25
%F	40	20	0	20	20
%ALL	35,2941176	11,7647059	5,8823529	23,5294118	23,5294118

Distribution of answers (percent values)

all	male	female	age: 20-29	age: 30 -39
-0,12	0,0	-0,4	-0,5	0,4

Mean values of answers to Question 6

Summary: Although both genders found the Socialite demonstrator interesting, innovative and exciting, female users gave significantly lower ratings than male users. Male users perceived the application more outstanding, exclusive and impressive while women rated it more second rate, standard and nondescript.

Also the age of the users seems to be an important discriminator. Younger users – between 20 – 29 years – gave more negative ratings (averagely -0,4) while the elder group (between 30 – 39 years) rated the dialogues very positively (on average 1,14).

Another striking result is that the voting for facial expression and the gesture (movement) is more positive than it is for the voice, which was perceived as unnatural and not appropriate. A possible explanation is that people are less rigorous in their judgement of the animation as it is

particularly cartoonish (2d Flash animation), whereas the voice quality is very high given synthesized emotional speech. In other words, the voice is too 'normal' and thus not good enough compared to human voice. Evidence for this point of view is given in Schröder (2004)¹ where a study on speech perception shows that the emotionally rich synthesized speech employed in the Socialite demonstrator is perceived as superior to the standard, neutral speech.

4.4 eShowroom: Some General Data

- 11 questionnaires were collected, (created on 8 different clients).
- 241 log entries were made (representing the number of played animation).
- The evaluation period of the eShowroom was from 1.2.2004 until 25.5.2004, i.e., ~140 days.
- Most users who filled in a questionnaire (64%) stated to use animated characters on a regular basis, 9% rated themselves as an expert and 18% told, that this was the first time meeting an animated avatar, 9% gave no answer.
- Most users (64%) stated to have 'expert knowledge' about computers, and 18% had 'no knowledge' and another 18% gave no answer to this question.
- 73% of the users were males and 18% females, 9% gave no answer.
- 55% were between 30 and 39 years, 18% were younger than 19, 9% were between 40-49, another 9% were between 20 and 29, and another 9% gave no answer.

¹ Schröder, M. (2004) Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. *PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.*

4.5 eShowroom: Evaluation of the Questions

The evaluation is analogous to the Socialite evaluation. In the following, we only show the mean values as this information is sufficient to illustrate the tendencies. As regards the age groups employed for cross-classification in the Socialite evaluation, respective data are not available for the eShowroom.

- Question 1 (Overall enjoy):

all	male	female
0,27	0,75	-1

Mean values of answers to Question 1

- Question 2(Movement):

all	male	female
0,0	0,25	0

Mean values of answers to Question 2

- Question 3(Voice):

all	male	female
-0,55	-1,0	0

Mean values of answers to Question 3

- Question 4:

	all	male	female
Hervorragend - Zweitklassig	-0,73	-0,38	-2
Exklusive - Standard	-0,6	-0,25	-2
Eindrucksvoll-Unbestimmbar	-0,18	-0,125	-1

Einzigartig - Gewöhnlich	-0,1	0	-1
Innovativ - Konservativ	0,36	0,13	1
Aufregend - Platt	-0,9	-0,5	-2
Interessant - Langweilig	-0,8	-0,38	-2

Mean values of answers to Question 4

- Question 5 (Tina performed)

all	Male	female
0,8	0,75	1

Mean values of answers to Question 5

- Question 6 (Ritchie performed.):

all	Male	female
0,36	0,25	2

Mean values of answers to Question 6

Summary: Female users gave on average significantly lower ratings than male users, and did not enjoy the dialogues in contrast to male users. Both genders perceived the application not outstanding, exclusive, interesting, exciting and impressive but a bit innovative. For these questions the tendency of both genders was the same, but male users gave more positive responses whereas the judgements of females were negative except for innovation. Possibly this negative attitude of females towards the application is an artefact of the car sales domain. The general attitude towards the application as a whole was contrasted by a positive opinion of the users concerning the performance of the avatars– which was ranked especially high by women users. While male users were more positive to the female character Tina, female users were more positive (unanimously highest rating) towards the male character Ritchie. However, as the average rating over male and female users is concerned, Tina's 'overall performance' is rated higher than Ritchie's (mean value .8 versus .36).

Again the voice (a high quality state of the art unit selection approach) lost over the animation as it was perceived as unnatural and not appropriate with the male users showing a clearly negative judgement whereas the female users were neutral, i.e., found it neither good nor bad.

5. Conclusion

We have presented the results of a questionnaire-based field test for the NECA demonstrators eShowroom and Socialite. Due to the small number of data, the findings need to be understood as first impressions on user perception of web applications featuring animated conversation among embodied conversational characters. These findings are valuable information in a user centered design approach.

We have learned that such applications like Socialite and eShowroom have a potential to attract people, see the high return rate to the Socialite demonstrator, and the number of presentation scenes generated in eShowroom. Whereas Socialite was positively attributed with innovation and excitement, the innovative aspect prevailed in the eShowroom. In general, males were more enthusiastic about the applications than were females. However, especially in the eShowroom the female users had a clearly positive opinion of the performance of the presentation agents, particularly of the male character.

Generally, the users were very critical about the applications and animations, with the eShowroom getting lower ratings than Socialite. This may be partially due to the domain (car sales is not particularly thrilling), but also partially due to the animation. The eShowroom with its 3d animation comes closer to TV commercials and high end Hollywood productions such as Shreck, and thus creates higher expectations about the animations which cannot be met and are not in focus in a 2 1/2 year R&D project concentrating on full computer generation of affective animated dialogue without any manual fine-tuning. In contrast to the eShowroom, Socialite, especially in combination with 'derSpittelberg', offers a more intriguing domain, i.e., the users can get in contact with each other, indirectly via their avatar and directly via e-mail and chat. Even though both the Socialite and the eShowroom animations are cartoonish, the Socialite 2d Flash animations are more crude than the eShowroom 3d animations, but get much better ratings than the latter (see question 2). We count this as further evidence for findings that have been made elsewhere that the closer the artificial world comes to the real world the more critical the user become.

A special issue is speech. Here humans are blistering. Where there is a bandwidth of acceptance for animation styles from LaLinea to Final-Fanatomy, there seems to be a much stronger positive/negative approach to speech by the nonexpert user. However, this is clearly relativised in studies comparing different approaches to speech synthesis, cf. Schröder (2004). Accordingly, we also found in our data that users were less critical with animation than with speech, although speech synthesis has strongly improved in the past years, and the voices employed in the NECA demonstrators are high end -- in the case of eShowroom, a high quality unit selection approach to synthesis is used; for Socialite we have specifically developed an approach to concatenative synthesis of affective speech including novel databases of expressive voice.

Summing up, customization of mulitmodal presentations/applications to the task at hand and the user group of interest seems crucial for the success of such an application. With the NECA platform we have made available a customizable and flexible tool for the design and implementation of applications featuring animated conversation.