Emo Chatting

Brigitte Krenn, Stephanie Schreitter

Executive summary

While in deliverable E3_1, the user evaluations of the four Bartender systems were presented by way of aggregating the Likert-scale ratings per system and male and female user group. The goal was, to get a first impression of potential areas of differences between female and male participants.

During the second stage of our experimentation process, we conducted a Mann-Whitney-U test on the data, in order to test for significances of differences between male and female evaluations of the systems. As already expected from the first experimentation phase, statistically significant differences between male and female users were rare in the specific data set.

In particular, a statistically significant difference between female and male evaluation could only be found for the autonomous conversational system displaying facial expression (system AB2) and only with respect to enjoyment (question: *How did you enjoy chatting with the Virtual Humans?*). **Females enjoyed chatting with autonomous conversational system displaying facial expression significantly more than males did.**

Moreover, we also reanalyzed the dialogues according to differences in the dialogue acts used and the appearance of LIWC categories using a Mann-Whitney-U test. Significant differences were also rare for dialogue acts and LIWC categories.

As regards dialogue acts, only in two cases, significant differences could be found:

- During the conversation in the **WOZ** scenario without facial expressions **female** users asked significantly more **`whQuestions'**.
- **Female** participants also showed significantly more **`Emphasis'** while chatting with the **conversational agent with facial expressions**.

LIWC categories: When communicating with the **autonomous agent displaying facial expression, females** used significantly more words from the LIWC categories **`cognitive mechanisms'**, whereas **men** used significantly more words from the category **`leisure'**.

To strengthen these evidences, data from larger user numbers would be necessary. We could not obtain further data under this specific experimental setting, mainly because the project CyberEmotions did not carry on experimentation using the specific scenario, and OFAI had no access to the 3D-engine and the WoZ system employed in the experiment.

As an alternative and extension of perspective, the assessment of male and female differences in affective dialogue settings with an artificial dialogue partner was carried on based on an

autonomous dialog system equipped with three affective profiles: a positive, a negative, and a neutral one. Results from this experiment are presented in E3_2_2.

In the following, we briefly recall the experimental setting, and then present the results after applying the Mann-Whitney-U test to the Bartender data.

Experimental setting

The data analysed in the following section was collected within the Cyber Emotions project. The experiment was carried out in a virtual bartender setting, in which an artificial bartender engaged a human user in a conversation which influenced the human communication partner either positively or negatively.

The virtual bartender (Figure 1) is the conversational front-end either to an autonomous conversational agent or to a Wizard-of-Oz (WOZ) system with the goal to engage the user in affective conversation. Each participant communicated with two versions of each system: with and without facial display of emotions for the agent and the user avatar.

Following each conversation, the users had to answer three questions:

- *How did you enjoy chatting with the Virtual Humans?*
- *Did you find a kind of `emotional connection' between yourself and the virtual human?*
- *Did you find the dialog with the virtual humans to be realistic?*

The answers ranged on a Likert-scale from 1 (*not at all*) to 6 (*very much*).

Figure 1: Bartender scenario (CyberEmotions)

Data analysis

In the analysis, we investigated whether male and female users had different impressions of the four conversational systems. 35 participants took part in all four experimental settings.

In Fig. 1, the mean values of the male (22) and the female (13) users are plotted. The participant's age ranged from 18 to 45 and 60% were younger than 30 years.

A Mann-Whitney-U test was conducted to test statistical significance of evaluation differences between female and male users. It is a non-parametric statistical hypothesis test for assessing whether two independent samples of observations, in our case the observations from the male and female users, have equally large values. The test is one of the standard non-parametric significance tests. It is more robust than for instance the Student's t-test and it is suitable for ordinal data, i.e. when we cannot assume that the spacing between adjacent values is uniform, which is very hard to argue for in the case of Likert-scale data.

As can be seen in Table 1, there is only a significant difference in the evaluation between male and female participants regarding the enjoyment of the virtual conversational system with facial expressions, i.e. female user enjoyed the conversation with the virtual conversational system with facial expressions significantly more than male users.

Figure 2 Line diagramm of mean scores for male and female users. AB refers to the autonomous conversational system, WOZ to the WOZ-system, 1 refers to a system without facial expressions and 2 to a system with facial expressions.

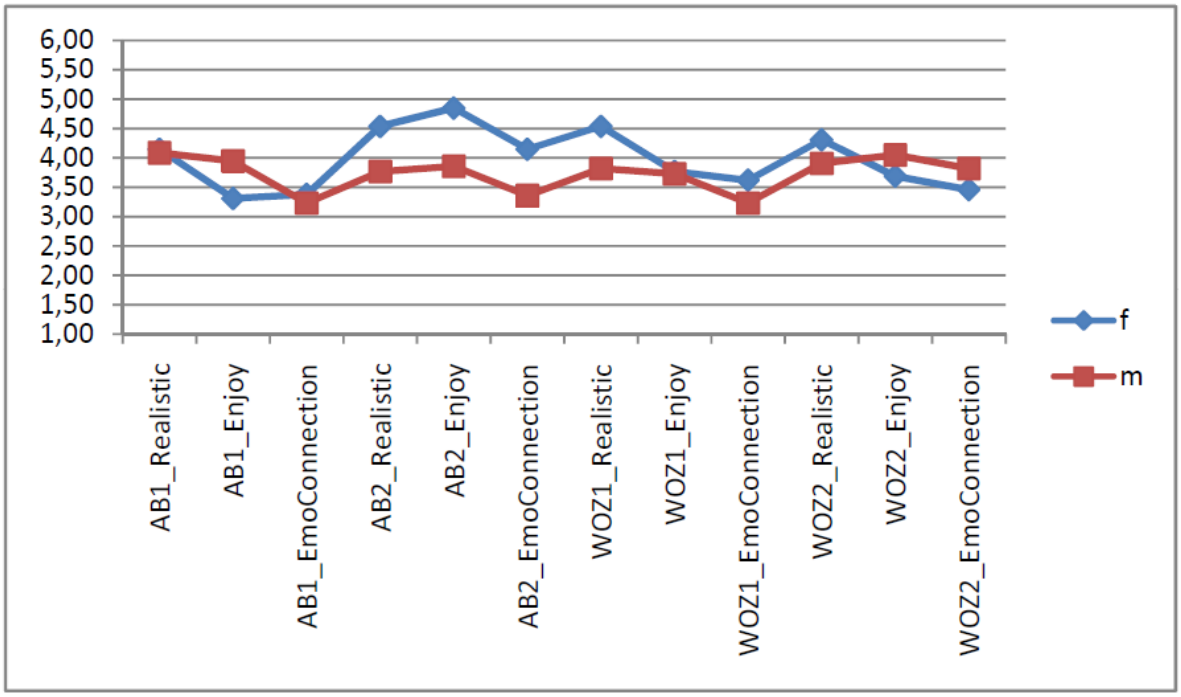

Table 1:The mean scores per question and system and the p-values (grey indicates significance) are listed.

	Mean values		P-values
	f	m	
AB1_Realistic	4.15	4.09	0.724
AB1_Enjoy	3.31	3.95	0.257
AB1_EmoConnection	3.38	3.23	0.699
AB2_Realistic	4.54	3.77	0.130
AB2_Enjoy	4.85	3.86	0.029
AB2_EmoConnection	4.15	3.36	0.106
WOZ1_Realistic	4.54	3.82	0.130
WOZ1_Enjoy	3.77	3.73	0.933
WOZ1_EmoConnection	3.62	3.23	0.335
WOZ2_Realistic	4.31	3.91	0.448
WOZ2_Enjoy	3.69	4.05	0.448
WOZ2_EmoConnection	3.46	3.82	0.389

Dialog Analysis

A Mann-Whitney-U test was applied to the data to investigate differences between to investigate differences in male and female dialog acts. It turned out that the chatting behaviour shown by female and male participants was very similar regarding dialog acts. Only in two cases, significant differences could be found:

- During the conversation in the **WOZ** scenario without facial expressions **female** users asked significantly more **`whQuestions'**.
- **Female** participants also showed significantly more **`Emphasis'** while chatting with the **conversational agent with facial expressions**.

Dialog Acts

	AB 1			WOZ 1			AB 2			WOZ 2		
	mean value		p value	mean value		p value	mean value		p value	mean value		p value
	female	male		female	male		female	male		female	male	
Statement	0.601	0.648	0.180	0.512	0.501	0.649	0.550	0.554	0.933	0.533	0.587	0.649
whQuestion	0.085	0.064	0.533	0.030	0.126	0.045	0.062	0.145	0.079	0.053	0.078	0.775
Reject	0.036	0.026	0.699	0.030	0.008	0.389	0.051	0.028	0.489	0.029	0.006	0.601
Emotion	0.005	0.000	0.724	0.000	0.032	0.511	0.000	0.000	1.000	0.038	0.006	0.578
Accept	0.029	0.031	1.000	0.019	0.041	0.511	0.016	0.024	0.827	0.010	0.034	0.468
Order	0.063	0.089	0.601	0.139	0.077	0.880	0.078	0.103	0.625	0.124	0.106	0.578
ynQuestion	0.092	0.062	0.335	0.118	0.122	0.906	0.070	0.079	0.625	0.112	0.058	0.749
Continuer	0.000	0.000	1.000	0.028	0.009	0.601	0.000	0.009	0.827	0.000	0.000	1.000
Bye	0.024	0.018	1.000	0.000	0.000	1.000	0.008	0.011	0.906	0.000	0.000	1.000
Emphasis	0.022	0.019	0.674	0.020	0.012	0.775	0.090	0.006	0.016	0.013	0.006	0.880
Greet	0.009	0.010	0.960	0.075	0.060	0.724	0.000	0.000	1.000	0.074	0.043	0.428
yAnswer	0.035	0.019	0.649	0.030	0.005	0.578	0.067	0.016	0.169	0.015	0.040	0.649
nAnswer	0.000	0.013	0.674	0.000	0.008	0.827	0.010	0.025	0.933	0.000	0.036	0.389

Table 2: The mean scores of the analysis of the dialog acts and the p-values (grey indicates significance).}

LIWC

An analysis of the LIWC categories (see Table 3) also showed very few significant differences between male and female users.

When communicating with the **autonomous agent displaying facial expression, females** used significantly more words from the LIWC categories **`cognitive mechanisms'**, whereas **men** used significantly more words from the category **`leisure'**.

Table 3: The mean scores of the LIWC categories of male and female users and the p-values (grey indicates significance).

	LIWC											
	AB 1			WOZ 1			AB 2			WOZ 2		
	mean values		p-values	mean values		p-values	mean values		p-values	mean values		p-values
	female	male		female	male		female	male		female	male	
Words	44.538	38.727	0.468	28.923	29.000	0.801	37.615	37.636	0.827	31.846	31.273	0.775
Characters	212.538	182.182	0.389	141.154	132.273	0.853	174.923	177.682	0.960	148.615	142.091	0.880
Funct	0.185	0.187	0.468	0.168	0.183	0.139	0.192	0.184	0.335	0.178	0.183	0.649
Pronoun	0.059	0.060	0.906	0.059	0.059	0.960	0.059	0.056	0.674	0.055	0.059	0.724
Ppron	0.046	0.041	0.987	0.046	0.037	0.389	0.044	0.036	0.091	0.036	0.040	0.533
I	0.024	0.027	0.408	0.028	0.023	0.428	0.028	0.020	0.243	0.018	0.025	0.169
We	0.001	0.001	0.775	0.001	0.002	0.775	0.002	0.001	0.555	0.001	0.001	0.933
You	0.020	0.013	0.149	0.019	0.014	0.389	0.015	0.016	0.853	0.018	0.015	0.649
SheHe	0.001	0.001	0.775	0.001	0.000	0.880	0.000	0.001	0.649	0.001	0.001	1.000
They	0.000	0.001	0.933	0.001	0.000	0.880	0.001	0.001	0.853	0.001	0.001	1.000
Ipro	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000
Article	0.030	0.024	0.203	0.019	0.022	0.578	0.030	0.030	0.699	0.033	0.033	0.749
Verbs	0.064	0.061	0.933	0.064	0.066	0.601	0.060	0.060	0.933	0.056	0.065	0.121
AuxVb	0.040	0.035	0.468	0.042	0.043	0.749	0.032	0.036	0.243	0.036	0.040	0.448
Past	0.007	0.012	0.191	0.003	0.005	0.649	0.006	0.008	0.674	0.005	0.008	0.511
Present	0.054	0.045	0.389	0.056	0.055	0.699	0.053	0.051	0.801	0.048	0.055	0.353
Future	0.002	0.002	0.801	0.005	0.005	0.933	0.002	0.002	0.801	0.003	0.003	0.906
Adverbs	0.020	0.020	0.724	0.017	0.019	0.801	0.024	0.020	0.801	0.012	0.018	0.287
Prep	0.023	0.027	0.601	0.013	0.016	0.555	0.026	0.023	0.335	0.018	0.022	0.601
Conj	0.015	0.016	0.649	0.012	0.015	0.408	0.016	0.014	0.801	0.017	0.010	0.169
Negate	0.008	0.015	0.216	0.012	0.012	0.906	0.013	0.015	0.533	0.016	0.006	0.098
Quant	0.006	0.006	0.853	0.006	0.009	0.601	0.005	0.006	0.775	0.004	0.004	0.933
Numbers	0.000	0.002	0.775	0.004	0.004	0.674	0.000	0.003	0.353	0.002	0.007	0.353
Swear	0.001	0.000	0.601	0.001	0.004	0.775	0.001	0.001	0.853	0.001	0.001	0.827
Social	0.040	0.026	0.098	0.037	0.033	0.674	0.033	0.031	0.933	0.033	0.028	0.408
Family	0.000	0.002	0.578	0.001	0.000	0.880	0.001	0.001	0.960	0.001	0.001	1.000
Friends	0.002	0.000	0.371	0.001	0.000	0.880	0.003	0.003	0.853	0.001	0.002	0.801
Humans	0.003	0.001	0.216	0.001	0.002	0.801	0.001	0.004	0.180	0.003	0.002	0.906
Affect	0.045	0.053	0.649	0.053	0.053	0.801	0.041	0.048	0.674	0.062	0.055	0.389
Posemo	0.041	0.050	0.511	0.053	0.046	0.408	0.036	0.044	0.448	0.061	0.049	0.203
Negemo	0.004	0.003	0.408	0.001	0.007	0.319	0.006	0.005	0.960	0.002	0.006	0.319
Anx	0.001	0.001	0.827	0.001	0.000	0.601	0.001	0.001	0.827	0.002	0.001	0.853
Anger	0.001	0.001	0.775	0.001	0.004	0.775	0.002	0.001	0.578	0.001	0.001	0.827
Sad	0.000	0.000	0.906	0.001	0.001	0.960	0.000	0.002	0.353	0.001	0.003	0.489
CogMech	0.043	0.046	0.749	0.038	0.038	0.775	0.059	0.041	0.041	0.041	0.032	0.191
Insight	0.008	0.005	0.555	0.003	0.004	0.853	0.002	0.003	1.000	0.003	0.003	0.960
Cause	0.002	0.006	0.287	0.003	0.004	0.775	0.004	0.006	1.000	0.004	0.003	0.555
Discrep	0.006	0.002	0.257	0.007	0.009	0.933	0.009	0.008	0.801	0.005	0.005	0.827
Tentat	0.007	0.007	0.987	0.010	0.011	0.880	0.006	0.010	0.203	0.008	0.006	0.555
Certain	0.003	0.003	0.749	0.004	0.001	0.159	0.008	0.002	0.389	0.003	0.007	0.601
Inhib	0.002	0.001	0.533	0.001	0.001	0.775	0.003	0.002	0.468	0.001	0.001	0.801
Incl	0.008	0.013	0.674	0.007	0.008	0.933	0.013	0.005	0.079	0.015	0.005	0.062
Excl	0.012	0.012	0.302	0.012	0.011	0.827	0.018	0.014	0.555	0.013	0.010	0.601
Percept	0.008	0.009	0.827	0.010	0.011	0.827	0.006	0.007	0.699	0.018	0.005	0.073
See	0.004	0.005	0.880	0.006	0.004	0.371	0.002	0.004	0.408	0.013	0.003	0.149
Hear	0.000	0.002	0.578	0.001	0.001	0.960	0.001	0.001	0.906	0.002	0.001	0.853
Feel	0.002	0.002	0.468	0.001	0.004	0.271	0.002	0.002	0.775	0.004	0.001	0.578
Bio	0.019	0.018	0.906	0.030	0.021	0.749	0.016	0.023	0.139	0.019	0.021	0.724
Body	0.001	0.001	0.775	0.002	0.003	0.511	0.001	0.002	0.699	0.001	0.001	0.960
Health	0.001	0.002	1.000	0.002	0.000	0.578	0.002	0.001	1.000	0.001	0.003	0.674
Sexual	0.000	0.001	0.933	0.001	0.002	0.775	0.001	0.001	0.533	0.001	0.001	0.933
Ingest	0.018	0.017	0.775	0.030	0.018	0.408	0.015	0.021	0.106	0.020	0.018	0.724
Relativ	0.024	0.022	0.533	0.019	0.017	0.987	0.020	0.024	0.511	0.018	0.024	0.933
Motion	0.004	0.004	0.625	0.001	0.000	0.880	0.006	0.003	0.257	0.002	0.005	0.448
Space	0.009	0.007	0.853	0.004	0.004	0.724	0.008	0.013	0.353	0.008	0.010	0.775
Time	0.014	0.017	0.827	0.020	0.017	0.801	0.010	0.014	0.389	0.016	0.012	0.389
Work	0.009	0.007	0.555	0.002	0.002	0.699	0.003	0.006	0.169	0.001	0.008	0.229
Achiev	0.003	0.002	0.853	0.005	0.005	0.724	0.002	0.001	0.749	0.003	0.006	0.448
Leisure	0.017	0.016	0.448	0.019	0.015	0.827	0.010	0.021	0.038	0.015	0.019	0.468
Home	0.001	0.000	0.601	0.001	0.000	0.880	0.001	0.002	0.601	0.001	0.001	1.000
Money	0.002	0.003	0.468	0.002	0.002	0.933	0.000	0.002	0.353	0.001	0.003	0.906
Relig	0.001	0.001	0.775	0.001	0.001	0.724	0.002	0.001	0.699	0.001	0.001	1.000
Death	0.000	0.001	0.933	0.001	0.000	0.880	0.000	0.001	0.827	0.001	0.002	0.801
Assent	0.010	0.014	0.408	0.011	0.012	0.906	0.013	0.009	0.468	0.007	0.011	0.625
Nonflu	0.002	0.001	0.775	0.002	0.005	0.533	0.001	0.001	0.852	0.001	0.001	1.000
Filler	0.001	0.005	0.649	0.007	0.005	0.749	0.004	0.004	0.906	0.007	0.002	0.319

Mann-Whitney-U versus Fisher's exact test

In the following, differences in the results of the Fisher's exact test (used in the first evaluation phase) and the Mann-Whitney-U test (used in the second evaluation phase) are outlined.

Although it is valid to define a p-value smaller than 0.1, in most studies results are said to be significant at the 0.05 level (see Agresti, Alan and Finlay, Barbara. Statistical Methods for the Social Sciences. Pearson Prentice Hall, New Jersey 2009). Therefore the results of the two test are compared for $p < 0.5$.

The Fisher's exact test as well as the Mann-Whitney-U test are significance tests for data containing two independent samples. While the Mann-Whitney-U test is a nonparametric test for comparing groups, the Fisher's exact test is applied especially to small samples and categorical data.

The dialog act analysis showed that male users asked significantly more whQuestions when chatting with the Wizard-of-Oz without facial expressions and female users showed significantly more emphasis towards the autonomous conversational system without facial expressions in both tests. Additionally, in the Fisher's exact test three more differences were significant: Male users asked significantly more whQuestions to both systems without facial expressions, while female users gave significantly more 'yAnswers' to the conversational system with facial expressions.

The analysis of the LIWC categories assessed two statistically significant differences for each test, all regarding the conversational agent with facial expressions. Men mentioned significantly more 'leisure' categories and significantly less 'inclusive' categories in the Fisher's exact test. Women mentioned significantly more 'cognitive mechanisms' in the Mann-Whitney-U test.

In spite of some differences, the two test show similar results.