Personal Report

Caroline Hallmann, 2640914

Communicative Robotics

Vrije Universiteit Amsterdam

Personal Report

Interactions with artificial social agents have started to become an important factor amongst today's society and have become a vital part in people's daily activities. The rapid developments of new technologies have caused a shift in the way humans interact with each other and their surroundings. Specifically, conversational AI is being increasingly implemented in various forms such as chatbots and social robots and other artificial social agents. Conversational AI research does not solely focus on one domain but rather requires the collaboration of experts from different scientific fields. Dialog system research is one of the underlying areas used to create conversational AI and artificial social agents (Deriu et al.,2021). When developing a conversational AI system, evaluation is a crucial component of the research process. Evaluation metrics are used to shine light on the best models and thus they strongly influence the research directions of a field (Howcroft et al.,2020). Artificial social agents are evaluated through 1) human/manual evaluation and 2) automatic evaluation. In this paper, the crucial components for evaluating artificial social agents will be discussed and analyzed from collected data of the Leolani experiment carried out by students in the communicative robot's course.

Manual evaluation

The first required step for conducting a manual evaluation of artificial social agents is setting up a robot interaction experiment. For the study, students engaged in two robot interaction experiments: 1) interaction1 and 2) interaction2. Experiment 1 focused on manual and automatic evaluation and experiment 2 only on automatic evaluation.

The 'interaction1' experiment involved the use of the Leolani chatbot and the local GraphDB server. 18 students chose to be one of the given fictional characters, and, interacted with the Leolani chatbot for 20 minutes, reaching a minimum of 100 turns. See Appendix 1 for the detailed description of experiment 1. My chosen character was Lenka and the conducted scenario had 120 turns. After all interactions were completed, a total of 18 scenario folders were uploaded under 'interaction1' including mine (Lenka), labelled as '9453e656-5f91-4bf6-9736-6fcd5ff3d532'.

Students performed manual evaluation on different scenario files by rating the conversation through the evaluation metrics for human evaluation from a scale of 0 (worst) to 5 (perfect). The evaluation metrics covered the following 9 components: 1) Overall Human Rating, 2) Interesting, 3) Engaging, 4) Specific, 5) Relevant, 6) Correct, 7) Semantically Appropriate, 8) Understandable and 9) Fluent.

Findings After conducting manual evaluation and analysing the different scenarios, several differences were found. First, not all interactions fulfilled the minimum requirement of 100 turns. Out of the 15 interactions, two did not reach 100 turns (Appendix 2, Table 1). Scenario Lenka '9453e' fulfilled the turn criteria with a total of 120 turns. The largest outlier is the score of 36 turns and is too low to be accounted for in the study. The second lowest turn score is 69 and lies in an abnormal distance from all the turn scores collected in the experiment. Thus, the two scenarios are not appropriate to use for the evaluation process as the low scores can cause tests to either miss significant findings or distort real results.

Evaluation metrics Each scenario was evaluated with the human evaluation metrics and scaled from 0 (worst) to 5 (perfect). For each scenario, the overall human rating, and the average score over the 8 submatrices were calculated.

1. *Overall human rating*

For each scenario, students submitted their overall human rating score for each turn. After generating all overall human rating scores of all given turns in the given scenario, the average of overall human rating was calculated (Appendix 2, Table 1). The highest human rating score can only be 5, and thus, the data representing a score of 22.02 is not a valid human rating score. As previously stated, two scenarios are to be disregarded due to their low number of turns. The data generated a human rating score of 0 for both scenarios and are also not valid. Scenario '7ed0b885-' has the highest average overall human rating score of 3.79. Majority of scenarios had a score between 3.50 and 2.50. Lenka scenario '9453e' has a score of 2.13, one of the lowest, followed by scenarios with score 1.95 and 0.

2. *Sub-metrics*

Each turn per scenario was manually evaluated through the given 8 sub-metrics and rated from 0 (worst) and 5 (perfect). After rating the 8 sub-metrics of all turns per scenario, the average of each sub-metrics was calculated and then the average over sub-metrics (Appendix 2, table 1). When comparing the average overall human rating and average over sub-metrics, several differences were found (Appendix 2, table 2). Table 2 (Appendix 2) shows that the average overall human rating is higher than the average over sub-metrics. However, as mentioned previously, the average overall human rating score of 22.02 cannot be accurate as it is above 5. The two scenarios with the lowest number of turns scored 0 in all sub-metrics and overall human rating. Overall, the average over sub-metrics is more accurate than the average overall human rating due to the lower numbers of outliers. Furthermore, average over sub-metrics requires more computation. After rating the 8 sub-metrics of all turns in a scenario, the average per sub-metrics is calculated. Then, the average over the 8 sub-metrics averages is taken. Overall human rating is evaluated by taking the average of the 8 sub-metrics per turn in a scenario. All turns receive an overall human rating score and then the average is calculated from those scores. Deriving overall human rating scores requires less computational power and has higher chances for human error compared to average over sub-metrics. Regardless, scenario '7ed0b885-' has the highest score in both metrics and the best human evaluation score out of all scenarios. For scenario Lenka '9453e', the average over sub-metrics is higher with a score of 2.71 compared to average overall human rating of 2.13. Overall, the scenario has a low human evaluation score compared to the rest of the population scores.

Automatic evaluations

The two main automatic evaluations used in this study were 1) Knowledge graphs and 2) USR. The automatic evaluations of robot interactions were carried out on experiment 1 (interaction1) and experiment 2 (interaction2). Experiment 1 required participants to take on the role of a fictional character and converse with the Leolani chatbot for min. 100 turns (See Appendix 1). For experiment 2, the Leolani software was implemented in the humanoid pepper robot and each participant engaged in a 10–15-minute verbal face-to-face robot interaction. See Appendix 3 for the detailed description of experiment 2.

Automatic evaluation: Knowledge graphs For interaction1, the automatic evaluations were conducted on 15 valid scenarios, including my individual scenario Lenka ‘9453e’. For the knowledge graph analysis, individual average degree and sparseness graphs were generated for each valid scenario. Two graphs, average degree (Fig.3), and sparseness (Fig.4) were generated on the data of scenario Lenka ‘9453e’.

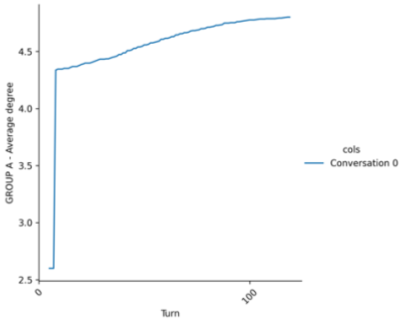


Fig.3 Average degree, scenario Lenka

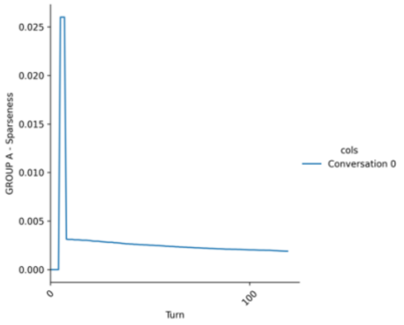


Fig.4 Sparseness, scenario Lenka

The average degree knowledge graph represents the fluency of the dialog. The flow of a conversation is captured through the average node degree by counting the incoming and outgoing edges per node in a knowledge graph (Báez Santamaría et al., 2022). The more fluent a conversation is, the lower the average degree over number of turns. In theory, human-human interactions are most fluent, and machine-machine interactions least fluent (Báez Santamaría et al., 2022). Thus, high fluency is represented through a steep downward slope and low fluency through a steep upward slope on the knowledge graph. Figure 3 shows the fluency of the Lenka scenario. The average node degree increases over time and is shown through the slope on the graph. The higher the average node degree, the steeper the upward slope, which shows that the scenario had low fluency. The conversation became less fluent towards the end of the conversation and may be due to repetitive nature of the Leolani chatbot. Previous studies found that the responses from Leolani are more repetitive than responses from other chatbots. Leolani becomes repetitive after several turns which then leads to an extreme increase of the average degree (ibid.).

Figure 4 shows the sparseness of scenario Lenka and represents the overall human rating of the conversation. Research states that a downward trend of sparseness per turn is linked to a more successful dialog as it indicates an increased semantic interconnectivity of the concepts in the dialog (ibid.). As seen in figure 4, the conversation starts at turn 0 and sparseness 0 since the knowledge graph does not contain any nodes yet, and consequently, cannot contain any edges. After the triple extractor found the first triple, sparseness increased to 0.026 but then dropped down to 0.004. Average degree (Fig.5) and sparseness (Fig.6) graphs containing all scenarios were then evaluated.

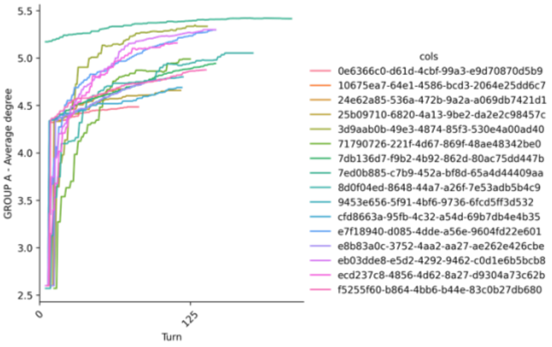


Fig.5. Average degree, all scenarios

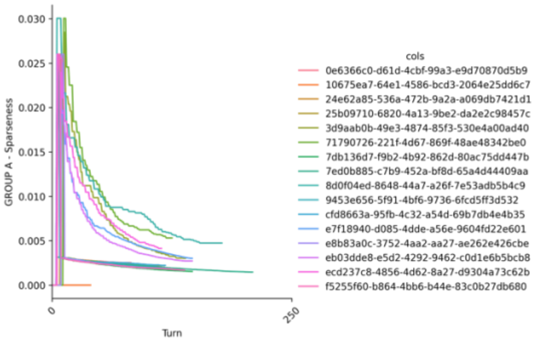


Fig.6. Sparseness, all scenarios

Figure 5 shows that all scenarios follow an upward slope and an increase in average degrees. Some scenarios have steeper upward slopes than others. The steeper the upward slope, the worse the fluency of the conversations. Overall, average degree scores ranged from 4.2 to 5.0. Scenario Lenka did not have the worst fluency of the population. Average degree of scenario Lenka matches the average scores of the population. The sparseness scores in figure 6 vary across scenarios. The green line peaking to 0.030 and ending at 0.005 had the best overall sparseness, thus, best overall human rating score. In comparison, scenario Lenka has one of the lowest sparseness scores (bottom 4) compared to the rest of the population.

Automatic evaluation: USR analysis Data from interaction1 and interaction2 was evaluated through the UnSupervised and Reference-free (USR) evaluation metric for dialog. The USR masked language modelling (MLM) metric uses a fine-tuned RoBERTa model to estimate the likelihood of a response (Mehri and Eskenazi., 2020). MLM likelihood evaluates the naturalness of responses in a conversation and how well the responses in a conversation were understood. The MLM likelihood (MLM_lhh) averages were calculated for interaction1 and interaction2. To evaluate interaction1, the MLM_lhh graph for scenario Lenka (Fig.7) and the graph with all scenarios (Fig.8) were used. For interaction2, the MLM_lhh graph of character Lenka was assessed (Fig 9).

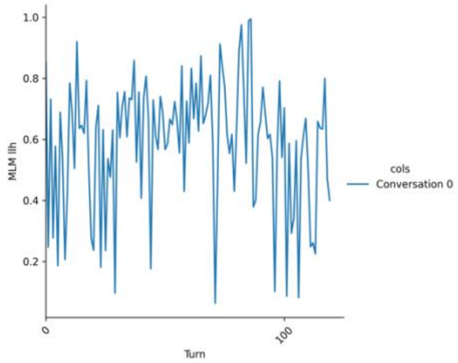


Fig.7. USR MLM likelihood, scenario Lenka, int1

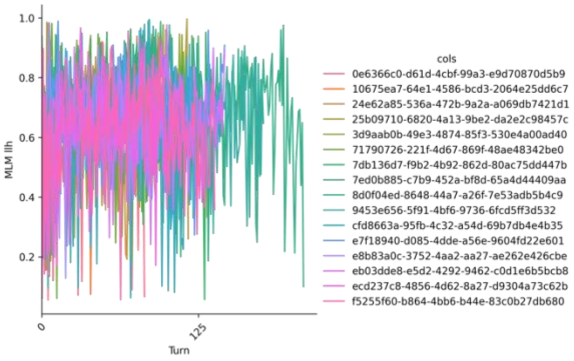


Fig.8. USR MLM likelihood, all scenarios, int1

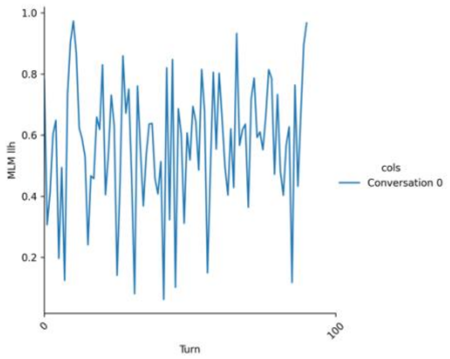


Fig.9. USR MLM likelihood, Lenka, int2

When comparing the two graphs of interaction1, the generated MLM_lhh scores of scenario Lenka align with the scores of the general population. The MLM_lhh scores closet to 1.0 in figure 7 are the most natural responses during the scenario. For scenario Lenka, responses were least natural at the start of conversation and became more natural over time, especially during the middle of conversation. However, at the end, the responses were not well understood. Overall, the conversation can't be defined as natural. Figure 8 shows that all scenarios had the lowest scores at the start of the conversation. Over time, scores increased, and the conversations become more natural. Ultimately, the scores of all scenarios fluctuate drastically (Fig.8) and thus, not considered natural. Interaction2 with Lenka (fig.9) has less turns than interaction1 with Lenka (fig.7). Figure 9 has greater fluctuations of MLM_lhh scores compared to figure 7. The results of interaction2 shows that less responses were rated at the start, indicating a possible delay of responses. After approx. 45 turns, more responses were evaluated and MLM_lhh scores started to increase, suggesting that the conversation became more natural over time.

Comparison across manual and automatic interaction

After conducting and interpreting the manual and automatic evaluations of my individual scenario 'Lenka' and total scenarios for interaction1 (See section manual evaluation and automatic evaluation) results did not correlate across the manual and automatic evaluations (Appendix 4). The manually generated overall human rating scores did not represent an accurate representation of findings and lacked validity due to the high number of outliers. Overall human rating scores were automatically evaluated through the sparseness knowledge graphs. The resulted downward trend of sparseness per turn for each scenario captured overall human ratings more accurate than the manual evaluation. The manual generated fluency scores (Appendix 2) are less accurate than the scores shown on the average node degree graph. The automatic evaluation is more appropriate as the flow of the conversation is represented via a slope across turns on the graph. The average of each sub-metrics in the manual evolution represents the overall score of each sub-metrics from total number of turns in a scenario. Average degree however analyzes each average degree score at each turn in the given scenario and the fluency of the conversation can traced back better to the given turn. Overall, the automatic evaluation generated more reliable scores than then manual evaluation.

Reflection on the chatbot and robot interactions

Interaction1 was evaluated manually and automatically, whereas interaction2 only automatically. The unimodal interaction with the chatbot took 20 minutes whereas the multimodal interaction with pepper lasted 10-15 minutes. As a result, interaction1 generated higher number of turns per scenario, making it more difficult to compare the results between the two interactions. Future research should apply all manual and automatic evaluations to both types of interaction and apply the same characters in both interactions. The time limit or min. total number of turns should be the same in both interactions as well.

References

Báez Santamaría, S., P. Vossen, and T. Baier. "Evaluating Agent Interactions Through Episodic Knowledge Graphs." *arXiv e-prints* (2022): arXiv-2209.

Howcroft, David M., Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. "Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions." In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 169-182. 2020.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review, 54(1):755–810

Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog GenerationLinks to an external site.. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Appendix 1

Description of Experiment 1: Interaction 1

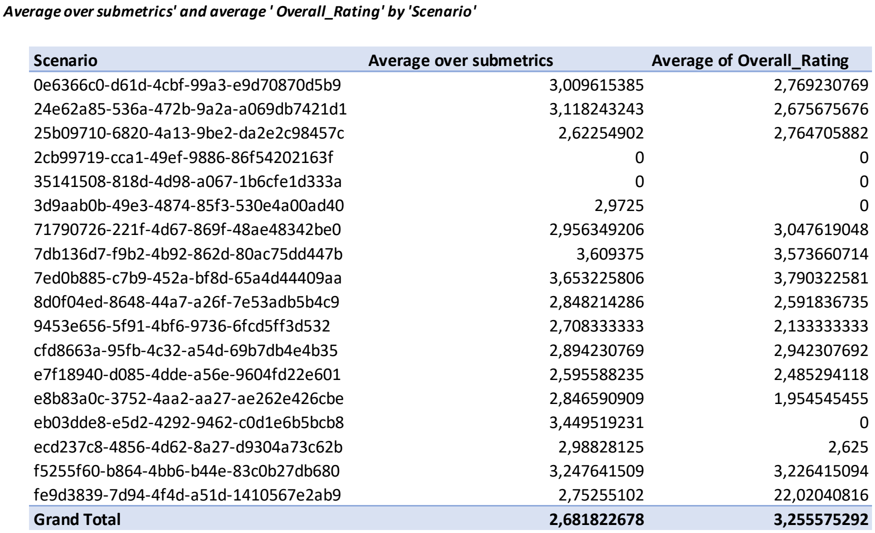
The first experiment involved the use of the Leolani chatbot and the local GraphDB server. A total of 18 students participated in the individual robot interaction experiment. First, students selected one of the given characters from the given list. For my individual experiment, the 'Lenka' character was chosen. The list included the character's name and some properties which they could use during the robot interaction. After selecting their character, participants had to set-up and run the Leolani chatbot on their computer. To set-up the Leolani chatbot, students had to create the clear repository "sandbox" in GraphDB and run Docker through the given instructions. Once the scenario was initialized, students had to visit a link which generated a chat UI for running the Leolani chatbot. To start the interaction, students had to type 'hello' to elicit Leolani's response ' Do you want to talk to me?' and ' What is your name?'. Students could only start chatting with Leolani after agreeing to talk with the agent and responding with their fake name. Once this was complete, students needed to interact with Leolani for 20 minutes and accumulate for at least 100 turns. The conversation was then terminated with the command 'bye' or 'goodbye' and then stored in the 'emissor' file. At the end, 18 scenario folders were uploaded under 'interaction1' including mine (Lenka), labelled as '9453e656-5f91-4bf6-9736-6fcd5ff3d532'.

Appendix 2

Table 1. Human evaluation scores ‘interaction1’

	Label	**7ed0b885-c7b9-452a-bf8d-65a4d44409aa**	**35141508-818d-4d98-a067-1b6cfe1d333a**	**25b09710-6820-4a13-9be2-da2e2c98457c**	**24e62a85-536a-472b-9a2a-a069db7421d1**	**2cb99719-cca1-49ef-9886-86f54202163f**	**e8b83a0c-3752-4aa2-aa27-ae262e426cbe**	**0e6366c0-d61d-4cbf-99a3-e9d70870d5b9**	**cfd8663a-95fb-4c32-a54d-69b7db4e4b35**	**ecd237c8-4856-4d62-8a27-d9304a73c62b**	**3d9aab0b-49e3-4874-85f3-530e4a00ad40**	**8d0f04ed-8648-44a7-a26f-7e53adb5b4c9**	**fe9d3839-7d94-4f4d-a51d-1410567e2ab9**	**f5255f60-b864-4bb6-b44e-83c0b27db680**	**eb03dde8-e5d2-4292-9462-c0d1e6b5bcb8**	**9453e656-5f91-4bf6-9736-6fcd5ff3d532**	**71790726-221f-4d67-869f-48ae48342be0**	**7db136d7-f9b2-4b92-862d-80ac75dd447b**	**e7f18940-d085-4dde-a56e-9604fd22e601**
0	Turns	210	36	118	102	69	101	83	119	115	140	178	145	139	147	120	126	147	147
1	Images	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Overall_Rating	3.790322581	0	2.764705882	2.675675676	0	1.954545455	2.769230769	2.942307692	2.625	0	2.591836735	22.02040816	3.226415094	0	2.133333333	3.047619048	3.573660714	2.485294118
3	Interesting	2.951612903	0	2.31372549	2.540540541	0	1.181818182	2.807692308	2.634615385	1.890625	1.18	2.091836735	1.714285714	3.056603774	2.865384615	2.066666667	2.777777778	2.910714286	1.941176471
4	Engaging	3	0	2.568627451	2.351351351	0	2.704545455	2.846153846	2.673076923	2.0625	2.22	2.102040816	1.795918367	3.716981132	3.903846154	2.1	2.349206349	3.321428571	1.955882353
5	Specific	3.532258065	0	2.901960784	3.189189189	0	2.659090909	3.153846154	3.134615385	2.546875	2.3	2.306122449	1.183673469	3.452830189	3.557692308	1.733333333	3.174603175	4.303571429	2.205882353
6	Relevant	4.096774194	0	2.607843137	2.810810811	0	2.954545455	2.961538462	2.884615385	2.953125	3.12	2.918367347	3.142857143	3.301886792	3.442307692	2.1	3.047619048	4.25	2.470588235
7	Correct	3.935483871	0	2.764705882	3.135135135	0	3.181818182	3.153846154	2.634615385	3.34375	2.7	2.826530612	2.857142857	2.79245283	3.384615385	2.9	2.968253968	3.803571429	3.073529412
8	Semantically_Appropriate	3.790322581	0	2.705882353	3.189189189	0	3.045454545	2.846153846	2.346153846	3.234375	4.1	3.112244898	4.040816327	3	3.230769231	3.533333333	3.126984127	3.071428571	2.779411765
9	Understandable	4.032258065	0	2.529411765	4.054054054	0	3.772727273	3.153846154	3.519230769	4.5	4.16	3.530612245	4.102040816	3.490566038	3.807692308	3.666666667	3.333333333	3.767857143	3.073529412
10	Fluent	3.887096774	0	2.588235294	3.675675676	0	3.272727273	3.153846154	3.326923077	3.375	4	3.897959184	3.183673469	3.169811321	3.403846154	3.566666667	2.873015873	3.446428571	3.264705882
	Average over submetrics	**3.653225806**	**0**	**2.62254902**	**3.118243243**	**0**	**2.846590909**	**3.009615385**	**2.894230769**	**2.98828125**	**2.9725**	**2.848214286**	**2.75255102**	**3.247641509**	**3.449519231**	**2.708333333**	**2.956349206**	**3.609375**	**2.595588235**

Table 2: submetrics VS overall rating



Average over submetrics' and average ' Overall_Rating' by 'Scenario'

Scenario	Average over submetrics	Average of Overall_Rating
0e6366c0-d61d-4cbf-99a3-e9d70870d5b9	3,009615385	2,769230769
24e62a85-536a-472b-9a2a-a069db7421d1	3,118243243	2,675675676
25b09710-6820-4a13-9be2-da2e2c98457c	2,62254902	2,764705882
2cb99719-cca1-49ef-9886-86f54202163f	0	0
35141508-818d-4d98-a067-1b6cfe1d333a	0	0
3d9aab0b-49e3-4874-85f3-530e4a00ad40	2,9725	0
71790726-221f-4d67-869f-48ae48342be0	2,956349206	3,047619048
7db136d7-f9b2-4b92-862d-80ac75dd447b	3,609375	3,573660714
7ed0b885-c7b9-452a-bf8d-65a4d44409aa	3,653225806	3,790322581
8d0f04ed-8648-44a7-a26f-7e53adb5b4c9	2,848214286	2,591836735
9453e656-5f91-4bf6-9736-6fcd5ff3d532	2,708333333	2,133333333
cfd8663a-95fb-4c32-a54d-69b7db4e4b35	2,894230769	2,942307692
e7f18940-d085-4dde-a56e-9604fd22e601	2,595588235	2,485294118
e8b83a0c-3752-4aa2-aa27-ae262e426cbe	2,846590909	1,954545455
eb03dde8-e5d2-4292-9462-c0d1e6b5bcb8	3,449519231	0
ecd237c8-4856-4d62-8a27-d9304a73c62b	2,98828125	2,625
f5255f60-b864-4bb6-b44e-83c0b27db680	3,247641509	3,226415094
fe9d3839-7d94-4f4d-a51d-1410567e2ab9	2,75255102	22,02040816
Grand Total	**2,681822678**	**3,255575292**

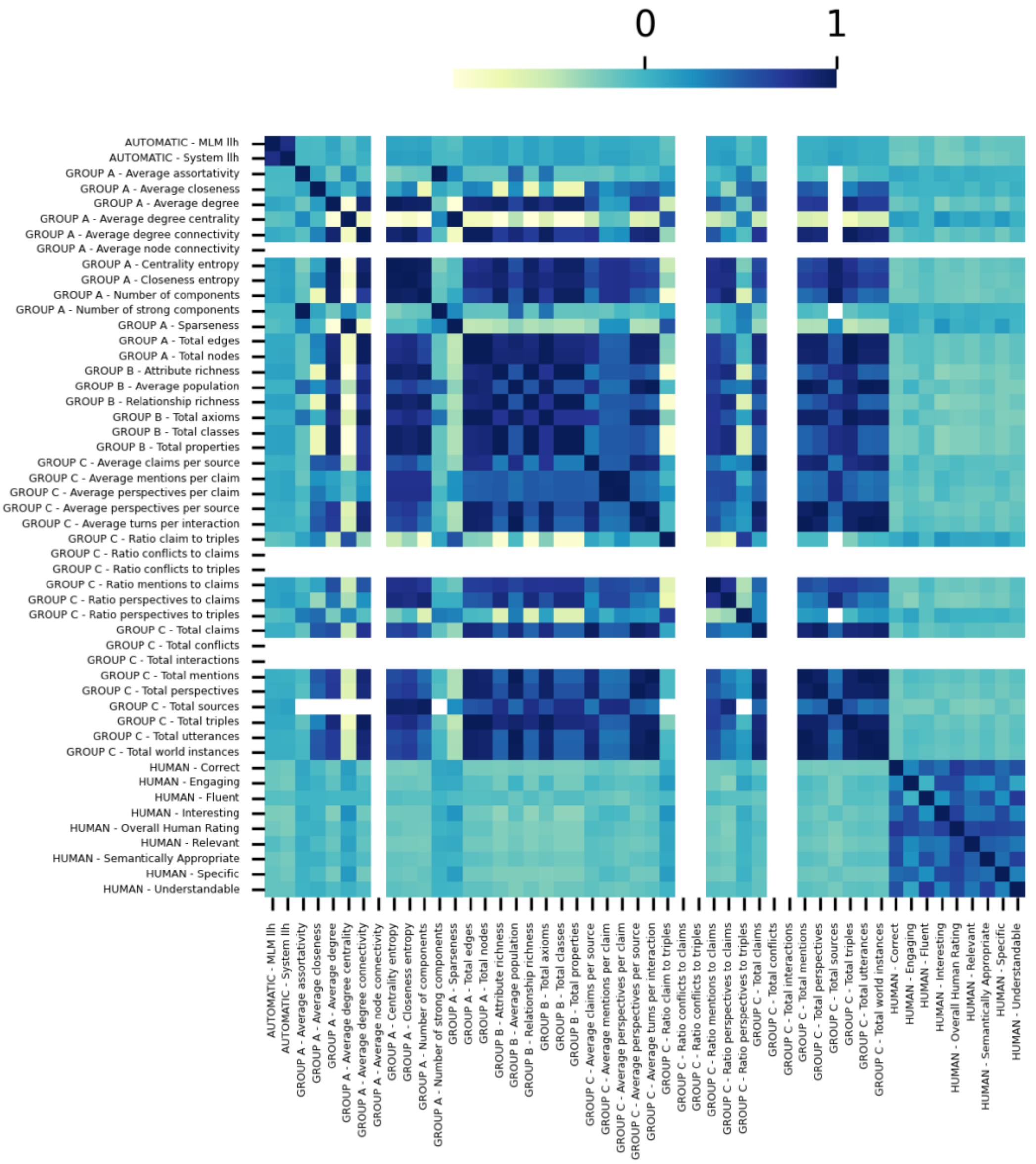
Appendix 3

Detailed description of Experiment 2: Interaction2

Human-robot interaction was captured through audio and visual data after students read and signed the consent form. The experiment took place in a room with three students presents where each had to interact with the humanoid pepper robot separately. A total of 15 students participated in the study. During the experiment, students had a 10–15-minute conversation with the pepper robot. The Leolani interface was implemented into pepper and the brain was merged in GraphDB. The experiment included placing several objects (bottle, teddy bear, cup, etc.) in view of the robot while students engaged in human-robot interaction. During the interaction, students chose to be a fictional character which the robot was familiar with and included references of other characters to elicit a flow like conversation. While engaging with the robot, students made references to visual context such as the objects in the room. To terminate the conversation, students used the command “goodbye”. A total of 15 characters were used during the robot interaction study. My character was named ‘Lenka’. At the end of each completed interaction, the emissor data (audio, image, text and rdf) was saved under a folder labelled with the corresponding character’s name. 15-character folders were saved under the main folder ‘interaction2’ and used for the automatic evaluations.

Appendix 4

Correlation heatmap: Automatic and manual evaluation


AUTOMATIC - MLM llh
AUTOMATIC - System llh
GROUP A - Average assortativity
GROUP A - Average closeness
GROUP A - Average degree
GROUP A - Average degree centrality
GROUP A - Average degree connectivity
GROUP A - Average node connectivity
GROUP A - Centrality entropy
GROUP A - Closeness entropy
GROUP A - Number of components
GROUP A - Number of strong components
GROUP A - Sparseness
GROUP A - Total edges
GROUP A - Total nodes
GROUP B - Attribute richness
GROUP B - Average population
GROUP B - Relationship richness
GROUP B - Total axioms
GROUP B - Total classes
GROUP B - Total properties
GROUP C - Average claims per source
GROUP C - Average mentions per claim
GROUP C - Average perspectives per claim
GROUP C - Average perspectives per source
GROUP C - Average turns per interaction
GROUP C - Ratio claim to triples
GROUP C - Ratio conflicts to claims
GROUP C - Ratio conflicts to triples
GROUP C - Ratio mentions to claims
GROUP C - Ratio perspectives to claims
GROUP C - Ratio perspectives to triples
GROUP C - Total claims
GROUP C - Total conflicts
GROUP C - Total interactions
GROUP C - Total mentions
GROUP C - Total perspectives
GROUP C - Total sources
GROUP C - Total triples
GROUP C - Total utterances
GROUP C - Total world instances
HUMAN - Correct
HUMAN - Engaging
HUMAN - Fluent
HUMAN - Interesting
HUMAN - Overall Human Rating
HUMAN - Relevant
HUMAN - Semantically Appropriate
HUMAN - Specific
HUMAN - Understandable
AUTOMATIC - MLM llh
AUTOMATIC - System llh
GROUP A - Average assortativity
GROUP A - Average closeness
GROUP A - Average degree
GROUP A - Average degree centrality
GROUP A - Average degree connectivity
GROUP A - Average node connectivity
GROUP A - Centrality entropy
GROUP A - Closeness entropy
GROUP A - Number of components
GROUP A - Number of strong components
GROUP A - Sparseness
GROUP A - Total edges
GROUP A - Total nodes
GROUP B - Attribute richness
GROUP B - Average population
GROUP B - Relationship richness
GROUP B - Total axioms
GROUP B - Total classes
GROUP B - Total properties
GROUP C - Average claims per source
GROUP C - Average mentions per claim
GROUP C - Average perspectives per claim
GROUP C - Average perspectives per source
GROUP C - Average turns per interaction
GROUP C - Ratio claim to triples
GROUP C - Ratio conflicts to claims
GROUP C - Ratio conflicts to triples
GROUP C - Ratio mentions to claims
GROUP C - Ratio perspectives to claims
GROUP C - Ratio perspectives to triples
GROUP C - Total claims
GROUP C - Total conflicts
GROUP C - Total interactions
GROUP C - Total mentions
GROUP C - Total perspectives
GROUP C - Total sources
GROUP C - Total triples
GROUP C - Total utterances
GROUP C - Total world instances
HUMAN - Correct
HUMAN - Engaging
HUMAN - Fluent
HUMAN - Interesting
HUMAN - Overall Human Rating
HUMAN - Relevant
HUMAN - Semantically Appropriate
HUMAN - Specific
HUMAN - Understandable