This has three levels: the no-explanation condition (*no-expl*); the preference-based explanations (*pref-expl*) condition, where the system uses the feature-based explanation obtained by the DM; and the social explanations condition (*soc-expl*), where the system uses the social explanation selected by the content planner.

5.2 Measurements

Participants were asked to answer two different questionnaires for this experiment: one measuring perceived quality of the conversational agent and the other measuring perceived quality of the overall interaction. For the former, we adapted the questionnaire used in [8]; this encompasses multiple aspects of a recommendation system's task performance and thus helped us analyze the potential trade-offs between the various independent variable conditions. The eight different items we used to measure task performance are listed in Table 3.

For the quality of the interaction, we relied on *rapport* [31], a notion commonly used in the domain of human-agent interactions to evaluate whether people are in sync with the system they are interacting with [11, 39]. The eight different items we used to measure rapport are listed in Table 4.

All answers for both questionnaires were on a 5-point Likert scale (anchors: 1 = completely disagree, 5 = completely agree).

5.3 Hypotheses

We hypothesized the following:

- **H1-a**: The type of recommendation (**Rec-Type**) delivered by the conversational agent will have a main effect on the agent's perceived quality. More specifically, the quality of the agent when delivering random recommendations (*rand-rec*) will be perceived as lower than the quality when delivering personalized recommendations (*pers-rec*).
- **H1-b**: The type of explanations (**Expl-Type**) used by the conversational agent will have a main effect on the agent's perceived quality. More specifically, the quality of the agent when using social explanations (*soc-expl*) will be perceived as higher than the quality when using preference-based explanations (*pref-expl*), which will in turn be perceived as higher than the quality when using no explanations at all (*no-expl*).
- **H2-a**: The type of recommendation (**Rec-Type**) delivered by the conversational agent will have a main effect on the perceived quality of the interaction. More specifically, the interactions with random recommendations (*rand-rec*) will be perceived as worse than the interactions with personalized recommendations (*pers-rec*).
- **H2-b**: The type of explanations (**Expl-Type**) used by the conversational agent will have a main effect on the perceived quality of the interaction. More specifically, interactions with social explanations (*soc-expl*) will be perceived as better than interactions with preference-based explanations (*pref-expl*), which will in turn be perceived as better than the interactions with no explanations at all (*no-expl*).

6 RESULTS

We collected 60 interactions, with 10 per condition. Our conversational agent recommended 140 movies in total, with an average of 2.33 recommendations per interaction (std = 1.5).

6.1 Quality of the conversational agent

We conducted a 2x3 factorial MANOVA (i.e., multivariate analysis of variance) with Rec-Type and Expl-Type as between-subject factors. The dependent measures were the eight questions presented in Table 3. The factorial MANOVA revealed two overall significant main effects of Rec-Type ($F(1, 54) = 6.3535$; $p < 0.0001$; Wilk's $\lambda = 0.48$) and Expl-Type ($F(2, 54) = 2.5508$; $p < 0.005$; Wilk's $\lambda 2 = 0.49$) on the perceived quality of the conversational agent. Both H1-a and H1-b are validated. The interaction between the two variables was not significant ($F(2; 54) = 0.689$; $p = 0.80$; Wilk's $\lambda = 0.80$).

Our follow-up analysis looked at univariate effects for each dependent measure with two-way ANOVAs and followed up with a post-hoc analysis when necessary. In Table 3, we report a summary of all means and standard errors (in parentheses) for the eight dependent variables. The differences between the means are marked according to their level of significance (* for $p < 0.05$, ** for $p < 0.005$ and *** for $p < 0.001$). We give more details about the follow-up analyses and discuss the results in the sections below.

6.1.1 Rec-Type vs. quality of the conversational agent. The type of recommendations delivered by the agent had a significant impact on the perceived quality of the system. Indeed, the results of the independent two-way ANOVAs showed a significant main effect of Rec-Type on all the dependent variables except for the perceived usefulness: decision confidence ($F(1, 54) = 48.672$; $p < 0.001$; $\eta^2 = 0.44$), user control ($F(1, 54) = 6.360$; $p < 0.05$; $\eta^2 = 0.09$), intention to return ($F(1, 54) = 9.371$; $p < 0.005$; $\eta^2 = 0.14$), perceived effort ($F(1, 54) = 17.184$; $p < 0.001$; $\eta^2 = 0.22$), intention to watch ($F(1, 54) = 28.767$; $p < 0.001$; $\eta^2 = 0.33$), recommendation quality ($F(1, 54) = 37.839$; $p < 0.001$; $\eta^2 = 0.35$), and transparency ($F(1, 54) = 4.382$; $p < 0.05$; $\eta^2 = 0.06$).

For all the questionnaire items, the agent was rated with higher scores when delivering personalized recommendations (*pers-rec*) than when delivering random ones (*rand-rec*).

6.1.2 Expl-Type vs. quality of the conversational agent. The results of the independent two-way ANOVAs showed a significant main effect of Expl-Type on four of the dependent variables: decision confidence ($F(2, 54) = 3.474$; $p < 0.05$; $\eta^2 = 0.06$), recommendation quality ($F(2, 54) = 7.703$; $p < 0.005$; $\eta^2 = 0.14$), perceived usefulness ($F(2, 54) = 6.677$; $p < 0.005$; $\eta^2 = 0.19$), and transparency:($F(2, 54) = 4.355$; $p < 0.05$; $\eta^2 = 0.12$)

The results of the post-hoc analyses (after Bonferroni correction) show that the agent was rated with a significantly higher score in decision confidence ($p < 0.05$) when using our model of social explanations (soc-expl) than when using preference-based explanations (pref-rec), and with a significantly higher score in recommendation quality ($p < 0.005$) and perceived usefulness ($p < 0.005$) compared to the two other levels of explanations (pref-expl and no-expl). The agent was rated with a significantly higher score ($p < 0.05$) when using preference-based explanations than when delivering recommendations without any explanation.

Dimensions	Subjective items	Rec-Type		Expl-Type		
		rand-rec	**pers-rec**	**no-expl**	**pref-expl**	**soc-expl**
Decision Confidence	The movies recommended to me during this interaction matched my interests.	2.07(±.90)***	3.80(±.77)***	2.90(±.97)	2.55(±.88)*	3.35(±.83)*
User Control	SARA allowed me to specify and change my preferences during the interaction.	3.23(±1.05)*	3.90(±.64)*	3.5(±.89)	3.2(±1.21)	4.0(±.71)
Intention to Return	I would use SARA to get movie recommendations in the future.	2.47(±1.34)**	3.50(±1.03)**	2.75(±1.06)	2.65(±1.27)	3.55(±1.35)
Perceived Effort	I easily found the movies I was looking for.	2.10(±1.04)***	3.33(±1.01)***	2.40(±.87)	2.55(±1.19)	3.20(±1.19)
Intention to Watch	I would watch the movies recommended to me, given the opportunity.	2.53(±1.22)***	3.07(±.53)***	3.20(±.88)	3.05(±1.06)	3.65(±1.02)
Recommendation Quality	I was satisfied with the movies recommended to me.	2.33(±1.05)***	3.93(±.57)***	2.85(±.82)**	2.70(±.91)**	3.85(±.97)**
Perceived Usefulness	SARA provided sufficient details about the movies recommended.	2.97(±1.14)	3.40(±1.01)	2.80(±1.15)**	2.80(±.98)**	3.95(±1.11)**
Transparency	SARA explained her reasoning behind the recommendations.	2.67(±1.37)*	3.40(±1.24)*	2.35(±1.34)*	3.60(±1.25)*	3.15(±1.25)

Table 3: Subjective questionnaire adapted from [8] to measure users' perceived quality of the system.

Dimensions	Subjective items	Rec-Type		Expl-Type		
		rand-rec	**pers-rec**	**no-expl**	**pref-expl**	**soc-expl**
Coordination	I felt I was in sync with SARA.	2.23(±.97)**	3.13(±1.04)**	2.25(±.85)*	2.60(±1.20)	3.20(±1.09)*
	I was able to say everything I wanted to say during the interaction.	3.13(±1.22)*	3.77(±1.00)*	3.05(±1.28)	3.70(±1.01)	3.60(±1.03)
Mutual Attentiveness	SARA was interested in what I was saying.	2.97(±1.14)*	3.70(±.66)*	3.35(±.98)	2.80(±1.07)*	3.85(±.93)*
	SARA was respectful to me and considered to my concerns.	3.30(±1.06)***	4.13(±.60)***	3.60(±.80)	3.50(±.82)	4.50(±.89)
Positivity	SARA was warm and caring.	3.10(±1.16)	3.47(±.89)	3.00(±.91)	3.25(±1.26)	3.60(±1.12)
	SARA was friendly to me.	4.10(±.74)	4.20(±.62)	3.95(±.80)	4.15(±.66)	4.35(±.67)
Rapport	SARA and I established rapport.	2.67(±1.10)*	3.27(±.80)*	2.75(±.86)	2.75(±1.28)	3.40(±1.03)
	I felt I had no connection with SARA.	3.50(±1.09)	2.90(±1.24)	3.50(±.86)	3.25(±1.39)	2.85(±1.33)

Table 4: Subjective questionnaire adapted from [39] to measure users' perceived quality of the interaction.

6.1.3 Discussion. As hypothesized, the type of recommendation had a significant impact on the perceived quality of the conversational agent. Participants were more satisfied with the agent when it delivered personalized recommendations matching their preferences, regardless of the type of explanation that was used.

Although the preference-based explanations helped participants better understand the reasoning behind the agent's recommendation, our model of social explanations helped them learn more details about the recommendation. One solution for solving this trade-off would be to combine these two types of explanations. Indeed, as noted in section 3.3, we noticed that humans often use feature-based explanations to link two successive recommendations before delivering more details on the current one (e.g., "Speaking of Tom Cruise movies, what about *Edge of Tomorrow*? I found it exceptionally well-made in every aspect, intriguing, exciting and even funny in the right way"). An alternative solution for the agent would be to frame its explanation negatively (e.g., "I don't like Tom Cruise, but I found *Edge of Tomorrow* exceptionally well-made in every aspect, intriguing, exciting and even funny in the right way"). However, expressing a disagreement towards one of the user's preferences might be harmful.

Unlike [17], we did not find any evidence that social explanations would increase users' intentions to return. However, participants who received social explanations were more satisfied with the recommendations and believed these recommendations better matched their preferences. These results show that a conversational agent able to give its "own" opinions and refer to its personal experiences is perceived as more convincing.

6.2 Quality of the interaction

We conducted a 2x3 factorial MANOVA with Rec-Type and Expl-Type as between-subjects factors. The dependent measures were the eight questions presented in Table 4. The factorial MANOVA revealed two overall significant main effects of Rec-Type ($F(1, 54) = 2.3955$; $p < 0.05$; Wilk's $\lambda = 0.71$) and Expl-Type ($F(2, 54) = 1.9608$; $p < 0.05$; Wilk's $\lambda 2 = 0.56$) on the perceived quality of the interaction. Both H2-a and H2-b are validated. The interaction between the two variables was not significant ($F(2; 54) = 1.2524$; $p = 0.24$; Wilk's $\lambda = 0.68$).

Similar to the previous section, we performed a follow-up analysis that looked at univariate effects for each dependent measure with two-way ANOVAs and followed with a post-hoc analysis when necessary. In Table 4, we report a summary of all means and standard errors (in parentheses) for the eight dependent variables. The differences between the means are marked according to their level of significance (* for $p < 0.05$, ** for $p < 0.005$ and *** for $p < 0.001$). We give more details about the follow-up analyses and discuss the results in the sections below.

6.2.1 Rec-Type vs. quality of the interaction. The results of the independent two-way ANOVAs showed a significant main effect of Rec-Type on five dependent variables: the two items measuring coordination ("I felt I was in sync with SARA" ($F(1, 54) = 9.663$; $p < 0.005$; $\eta^2 = 0.13$) and "I was able to say everything I wanted during the interaction" ($F(1, 54) = 4.292$; $p < 0.05$; $\eta^2 = 0.07$)), the two items measuring mutual attentiveness ("SARA was interested in what I was saying" ($F(1, 54) = 6.892$; $p < 0.05$; $\eta^2 = 0.10$) and "SARA was respectful to me and considered to my concerns" ($F(1, 54) = 12.255$; $p < 0.001$; $\eta^2 = 0.17$)), and one item measuring rapport ("SARA and I established rapport" ($F(1, 54) = 4.154$; $p < 0.05$; $\eta^2 = 0.06$)).

In all these cases, the agent was rated with higher scores when delivering personalized recommendations (*pers-rec*) than when delivering random ones (*rand-rec*).

6.2.2 Expl-Type vs. quality of the interaction. The results of the independent two-way ANOVAs showed a significant main effect of Expl-Type on two dependent variables: one item measuring coordination ("I felt I was in sync with SARA" ($F(2, 54) = 3.474$; $p < 0.05$; $\eta^2 = 0.10$)) and one measuring mutual attentiveness ("SARA

was interested in what I was saying" ($F(2, 54) = 4.714$; $p < 0.05$; $\eta^2 = 0.13$)).

The results of the post-hoc analyses (after Bonferroni correction) show that the agent was rated with a significantly higher score in the coordination item ($p < 0.05$) when using our model of social explanations (soc-expl) compared to when using no explanation (no-rec), and with a significantly higher score in the mutual attentiveness item ($p < 0.005$) compared to the pref-expl condition.

6.2.3 Discussion. Participants preferred interacting with a conversational agent delivering personalized recommendations. This result matches with the findings from [12], which explains that while the interview phase might help find more relevant items for users, the additional questioning might lead to disappointment if the recommendation does not meet the user's expectations. This also shows that in a recommendation context, a conversational agent's task-performance influences rapport through enhanced coordination and mutual attentiveness, regardless of the explanations it uses.

Regarding the explanations, participants felt they were more in sync with a conversational agent using social explanations and considered the agent as more interested in what participants were saying. This can be linked to the computational model of rapport proposed in [38]: disclosing topic related personal information improves both mutual attentiveness and coordination. This is also consistent with our above results showing that participants who received social explanations found their recommendations to be more relevant; participants felt that the conversational agent was more interested in what they were saying, which resulted in a better-informed recommendation.

7 CONCLUSION

In this paper, we presented the human-centered design implemented in our conversational recommendation agent. Our model of social explanations, constructed through careful annotation and analysis of a relevant corpus, leveraged observed probabilities for identified categories and subcategories of recommendations. This was incorporated in the form of a content planner within our conversational agent's architecture. Our user experiment evaluated the influence of these social explanations on the perceived quality of our system as well as the interaction; results indicate that they significantly improved both. Moreover, a system using social explanations was perceived as more in sync with its users and more interested in what they were saying. This aligns with [17] and emphasizes the need to endow conversational recommendation systems with social conversational strategies, as well as to build systems able to express personal opinions and experiences.

One potential extension of this work would be to overcome the limited size of our initial corpus by annotating a larger dataset of movie reviews using our explanation categories. That would allow us to refine our content planner and would provide us with more examples to generate natural sentences. Although endowing our agent with a human-like identity might seem inappropriate (e.g. users know the agent cannot watch movies in theaters), the results from [9] show that the type of identity revealed by a virtual character (human-like vs. artificial) does not influence people's perception.

Another way to improve the perceived quality of the system and/or interaction would be to optimize the interview phase as suggested by [15]. Although almost all of the participants had a preferred movie genre, only a few specified a favorite director. Soliciting too many specific preferences could be stress-inducing, and participants might consequently overthink their responses. Moreover, as described in [23], a conversational recommendation system using sentences that are too-long (compared to the user's utterances) will decrease the quality of the interaction, and the recommendations are less likely to be approved. We thus seek to extend our sentence planner such that it can adapt the length of its explanations based on the length of the user's sentences; this improved means of generation will likely result in "better" recommendations.

ACKNOWLEDGMENTS

This work was supported in part by funding from Oath and the IT R&D program of MSIP/IITP [2017-0-00255, Autonomous digital companion development]. We would also like to thank John Choi for his generous help, Yoo Jin Shin for refining annotations, and the members of Carnegie Mellon University's ArticuLab for their feedback and support.

REFERENCES

[1] Amos Azaria and Jason Hong. 2016. Recommender systems with personality. In *Proceedings of the 10th ACM conference on Recommender systems*. ACM, 207–210.
[2] J. Berant, A. Chou, R. Frostig, and P. Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*.
[3] George Y Bizer, Jeff T Larsen, and Richard E Petty. 2011. Exploring the valence-framing effect: Negative framing enhances attitude strength. *Political psychology* 32, 1 (2011), 59–80.
[4] Giuseppe Carenini, Jocelyin Smith, and David Poole. 2003. Towards more conversational and collaborative recommender systems. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 12–18.
[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.
[6] Rose Catherine and William Cohen. 2016. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 325–332.
[7] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 175–182.
[8] Li Chen and Pearl Pu. 2008. A cross-cultural user evaluation of product recommender interfaces. In *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 75–82.
[9] Setareh Nasihati Gilani, Kraig Sheetz, Gale Lucas, and David Traum. 2016. What kind of stories should a virtual human swap?. In *International Conference on Intelligent Virtual Agents*. Springer, 128–140.
[10] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, 517–520.
[11] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. 2006. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*. Springer, 14–27.
[12] Ulrike Gretzel and Daniel R Fesenmaier. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce* 11, 2 (2006), 81–100.
[13] Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 53–60.
[14] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1661–1670.

[15] Michael Jugovac and Dietmar Jannach. 2017. Interacting with recommenders—overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10.

[16] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[17] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[18] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[19] Theodora Nanou, George Lekakos, and Konstantinos Fouskas. 2010. The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. *Multimedia systems* 16, 4-5 (2010), 219–230.

[20] Neal R Norrick. 2005. Interactional remembering in conversational narrative. *Journal of Pragmatics* 37, 11 (2005), 1819–1844.

[21] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[22] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583.

[23] Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field Trial Analysis of Socially Aware Robot Assistant. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1241–1249.

[24] Ivens Portugal, Paulo Alencar, and Donald Cowan. 2018. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications* 97 (2018), 205–227.

[25] Arpit Rana and Derek Bridge. 2018. Explanations that are Intrinsic to Recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 187–195.

[26] Derek D Rucker, Richard E Petty, and Pablo Briñol. 2008. What's in a frame anyway?: A meta-cognitive analysis of the impact of one versus two sided message framing on attitude certainty. *Journal of Consumer Psychology* 18, 2 (2008), 137–149.

[27] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742* (2015).

[28] Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence* 19, 4 (2003), 311–381.

[29] Jan Svennevig. 2000. *Getting acquainted in conversation*. John Benjamins.

[30] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. MoviExplain: a recommender system with explanations. *RecSys* 9 (2009), 317–320.

[31] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1, 4 (1990), 285—-293.

[32] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[33] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[34] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Casell. 2018. Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *Proceedings of the 2018 International Workshop on Spoken Dialog System Technology*.

[35] Pontus Wärnestål, Lars Degerstedt, and Arne Jönsson. 2007. Interview and delivery: Dialogue strategies for conversational recommender systems. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*. 199–205.

[36] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 5.

[37] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).

[38] Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 514–527.

[39] Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. SOGO: A Social Intelligent Negotiation Dialogue System. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 239–246.

A ANNOTATION EXAMPLES

(a) Annotation Example 1

Speaker	Sentence	Annotation
A	So we went to see SOAPDISH and...	
B	Was it good?	
A	Oh, hysterical.	**[PO_POS]**
	We laughed so hard, it was just, you couldn't hear half the dialogue because everyone in the audience was laughing.	**[PE_A]**

(b) Annotation Example 2

Speaker	Sentence	Annotation
A	Have you seen the movie CLASS ACTION with Gene Hackman?	**[MF_C]**
B	No, I haven't yet.	
A	I saw it this weekend	**[PE_L]**
	and it is, uh, to me an outstanding movie.	**[PO_POS]**
	I thoroughly enjoyed it.	**[PO_POS]**
	He is, uh, an attorney and his daughter is an attorney and she has a suit against his company.	**[MF_P]**

(c) Annotation Example 3

Speaker	Sentence	Annotation
A	DANCES WITH WOLVES did not seem to have anything added. It was just a legitimate kind of film.	**[PO_ANA]**
	And that is the reason why I suppose it won so many Oscars,	**[MF_A]**
	because it really was good even though it is such a long movie.	**[PO_SO]**
	You know, they said, "Oh, people won't be interested in a three hours movie." But it certainly gotten good acclaim everywhere it has gone.	**[TPO_B]**

Figure 2: Three annotated movie chunks from the corpus.