

Discourse Behavior of Older Adults Interacting With a Dialogue Agent Competent in Multiple Topics

S. Zahra Razavi

University of Rochester
Rochester, NY, USA

srazavi@cs.rochester.edu

Lenhart K. Schubert

University of Rochester
Rochester, NY, USA

schubert@cs.rochester.edu

Kimberly A. Van Orden

University of Rochester Medical Center
Rochester, NY, USA

kimberly_vanorden@urmc.rochester.edu

Mohammad Rafayet Ali

University of Rochester
Rochester, NY, USA

mali7@cs.rochester.edu

Abstract

We present some results concerning the dialogue behavior and inferred sentiment of a group of older adults interacting with a computer-based avatar. Our avatar is unique in its ability to hold natural dialogues on a wide range of everyday topics—27 topics in three groups, developed with the help of gerontologists. The three groups vary in “degrees of intimacy”, and as such in degrees of difficulty for the user. Each participant interacted with the avatar for 7-9 sessions over a period of 3-4 weeks; analysis of the dialogues reveals correlations such as greater verbosity for more difficult topics, increasing verbosity with successive sessions, especially for more difficult topics, stronger sentiment on topics concerned with life goals rather than routine activities, and stronger self-disclosure for more intimate topics. In addition to their intrinsic interest, these results also reflect positively on the sophistication of our dialogue system.

1 Introduction

Spoken dialogue systems have proved beneficial for helping older people with their needs, including social companionship (Miehle et al., 2019; Abdollahi et al., 2017), health advice (Ono et al.), palliative care (Utami et al., 2017), reminiscence therapy (Arean et al., 1993) and many other applications. In many of these applications, it is crucial to keep users involved in the task for multiple sessions through engaging conversations. Here we seek insights into user behavior when interacting with a system capable of conversing about various casual topics.

Our study is based on linguistic data collected from 80 sessions of interaction with nine participants over a course of 3-4 weeks. The topics were suggested by gerontological experts and are categorized into three groups based on their degree of intimacy. We investigate how users’ verbosity, sentiment, and self-disclosure behavior depend on the topic under discussion and the avatar’s tone, and how they evolve with time. We measure verbosity in terms of users’ turn lengths, sentiment in terms of the Vader sentiment analysis tool, and self-disclosure in terms of cues suggested by the literature, and extracted using LIWC categories. We present the insights obtained from the analysis and discuss the results.

2 Background

Content analysis of dialogues with non-task oriented conversational agents (CAs) has proved helpful in increasing CAs’ effectiveness in a variety of tasks. For instance, detecting the main themes of a dialogue, or speech and language features, could assist in detecting schizophrenia (Del-lazizzo et al., 2018) and dementia (Ujiro et al., 2018), and preventing suicide (Martínez-Miranda, 2017). An important aspect of conversational content we investigate in this paper is the degree of self-disclosure. Encouraging self-disclosure can increase rapport in user-CA interaction (Pecune et al., 2018) and thereby the effectiveness of the virtual agent in different tasks (e.g., health coaching (Lisetti et al., 2013)). We also study the role of sentiment, since use of sentiment features has been observed to increase the quality of conversational agent output (Rinaldi et al., 2017). While

sentiment can be evaluated using analysis tools such as Vader (Hutto and Gilbert, 2014), the set of features indicative of self-disclosure remains ill-defined. (Ravichander and Black, 2018) suggested utterance length, negation words, POS tags, and emotion-laden words as self-disclosure markers in an open-ended conversation with a chatbot. (Houghton and Joinson, 2012) identifies personal pronouns, word count, and family and sexual words as significant, based on comparing secret tweets with normal tweets. (Bak et al., 2014) observed that tweets with deeper self-disclosure contain secretive wishes or sensitive information while medium self-disclosing tweets convey general information about self such as family, education, etc. In our work we tried several LIWC categories based on the cited literature.

3 Dataset

Our data came from multi-session interactions between a screen-based virtual agent and elderly users, where the agent leads users in casual conversations controlled by an automatic dialogue manager. The system was designed as a tool allowing users to practice their communication skills, giving them feedback on their non-verbal behavior and speech prosody. We recruited nine participants, each of whom had seven to nine sessions with the avatar; the first and the last interaction were held in the lab and the rest were self-initiated by users at home. Participants were asked to fill out surveys and were evaluated for their communication skills by experts.

Dialogues. Each interaction consists of 3 subsessions, each containing 3-5 questions from the avatar on a specific topic listed in table 1. The dialogue manager follows a plan for each topic, asks some questions, extracts essential information from users’ inputs and produces relevant comments indicating its understanding of the user. Each interaction took 15-20 minutes depending on the number of questions and the user’s verbosity.

The topics were selected by gerontological experts and divided into three groups based on their emotional intensity: easy, medium, and hard. “Easy” (less intimate) topics are ones likely to be broached in making someone’s acquaintance, while the harder ones are more emotionally evocative and call for more self-disclosure. As can be seen in table 1, the dialogue sessions were designed so that users start with easier topics in

	Subsessions Topics	EI
S1	Getting to know (I, II), Activity	E,E,E
S2	City you live in (I, II), Pets	E,E,M
S3	Family, Gathering, Yourself	E,M,H
S4	Weather, Driving, Cooking	E,H,E
S5	Outdoors, Travel, Plan for today	M,M,E
S6	Chores, Money, Growing older	E,M,H
S7	Education, Job, Life goals	M,M,H
S8	Technology, Books, Arts	M,M,M
S9	Sleep, Health, Exercise	M,M,M

Table 1: Dialogue topics and their emotional intensity level (E: easy, M: medium, H: hard)

Feature	Number
users interacted with the avatar	9
subsessions	198
total users’ turns	668
total users’ words	29054
total avatars’ words	24296

Table 2: Collected data statistics

earlier sessions and gradually transition to harder ones as they progress in the study.

Data statistics. We collected the transcripts (produced via ASR) from nine users interacting with the system over seven to nine days. A few subsessions were missed because of technical issues. Table 2 summarizes the collected data.

4 Dialogue Content Analysis

We analyzed three aspects of the dialogue content. The first concerns verbosity, where we looked for differences in verbosity across different sessions, users, and topic classes; we also analyzed changes in verbosity over time. The second concerns the results of sentiment analysis for different sessions and the tone change over time. The final aspect concerns the kinds of self-disclosure cues we gleaned from the literature.

4.1 Verbosity

Our metric for utterance length was word count.

Response length change over time. The results show that users on average tend to provide longer responses as they proceed in a conversation. Figure 1 shows average response length among all users in different subsessions. We also observe a strong, significant correlation (a) between the average word count and the particular subsession (Pearson $r = 0.76$, $p < 10^{-5}$); (b) between the

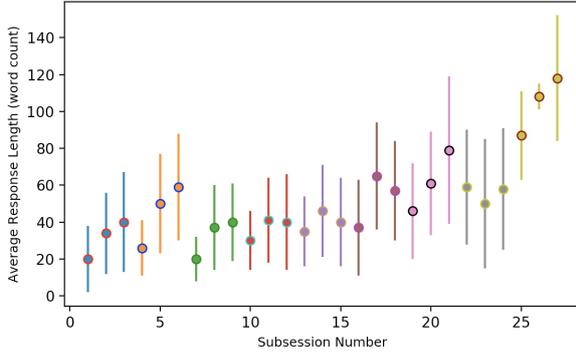


Figure 1: Users’ average turn length in each subsession

Topic class	$ave(Sent_a)$	$ave(Sent_u)$
Easy	0.3 ($\sigma = 0.31$)	0.43 ($\sigma = 0.35$)
Medium	0.36 ($\sigma = 0.3$)	0.62 ($\sigma = 0.31$)
Hard	0.38 ($\sigma = 0.32$)	0.63 ($\sigma = 0.3$)

Table 3: Average sentiment score for topic classes

average word count and the user’s turn number in the whole interaction ($r = 0.68$, $p < 10^{-12}$; and (c) between the average word count and the interaction number ($r = 0.81$, $p = 0.008$). Trends (b) and (c), however, are not the same for all individuals. For five out of nine users the turn length correlation with time is significantly strong, while for the rest we cannot see any significant correlation.

Users’ turn length and topic classes. The introduced topic classes significantly affect users’ the response length. The average among all users shows that users provide longer responses to “hard” questions, where the average is $57.60(\sigma = 33.92)$ words, while responses to “medium” and “easy” questions contain an average number of $51.41(\sigma = 30.39)$ words and $32.73(\sigma = 24.03)$ words respectively. Interestingly, the response length change over time is not significant for easy topics but it is significantly strong for medium ($r = 0.81$, $p < 10^{-3}$) and hard ($r = 0.94$, $p = 0.05$) topics.

User and avatar turn length. Some studies suggest that the utterance lengths of one speaker can influence the interlocutor’s utterance lengths. We looked for any correlation between the avatar’s input length and users’ corresponding turn length, but did not observe any meaningful relation.

4.2 Sentiment

We used VADER (Hutto and Gilbert, 2014) to quantify utterance sentiment for each avatar and

user turn.

User vs. avatar turn sentiment The correlation coefficient value shows a weak but significant correlation between a given user turn and the avatar’s preceding turn ($r = 0.23$, $p < 10^{-8}$); this suggests a slight dependence of the user’s sentiment on the avatar’s tone (though both might be derivative from the particular question content). In order to compensate for the possible influence of the avatar’s tone on the user, we study sentiment difference over time ($Sentiment_{user} - Sentiment_{avatar}$). We observe a significant weak increase in positive tone over time ($r = 0.35$, $p < 10^{-3}$).

Sentiment for different topics. A more careful look into different interaction sessions provides some insight into the relation between user sentiment and dialogue topics. We should first note that the avatar is designed to convey a positive, friendly tone in its interactions, thereby encouraging a generally positive tone on user’s side. However, we find user sentiment to be significantly more positive for some topics than others. Among them are “Travel”, with sentiment score= $0.74(\sigma = 0.22)$, “Health”, with sentiment score= $0.76(\sigma = 0.13)$, “Education”, with sentiment score= $0.74(\sigma = 0.07)$, and “Outdoor”, with sentiment score= $0.68(\sigma = 0.28)$. On the other hand, in talking about subjects such as “Family”, “Getting to know each other”, and “Managing money” people tend to be more neutral, with respective average sentiment scores of $0.045(\sigma = 0.18)$, $0.19(\sigma = 0.38)$, $0.32(\sigma = 0.39)$.

We infer that topics concerned with life goals evoke stronger emotions than those concerned with routine activities of daily life. As well, discussion of eventualities such as the death of a partner or living alone after others have moved out naturally leads to a more negative emotional tone. There are other themes that evoke both negative and positive user comments, and hence sentiment fluctuations resulting in a high standard deviation and no meaningful average. An example is the topic “Growing older” with sentiment score = $0.37(\sigma = 0.50)$.

Sentiment for different topic classes. We also studied the average sentiment for the three topic classes introduced in section 3. Our hypothesis was that emotionally evocative topics produce stronger user sentiment than more neutral ones.

Potential SD Cues	Easy	Medium	Hard
Word count per turn	31.97	49.61	55.45
1st person pron.	9.91	9.29	9.46
Family and friend	1.02	1.08	1.03
Negative emotion	0.51	0.54	0.84
Positive emotion	4.91	4.9	5.54
Drives	5.71	6.82	7.26
Personal concerns	5.2	7.07	4.15

Table 4: LIWC score of SD cues for three topic classes

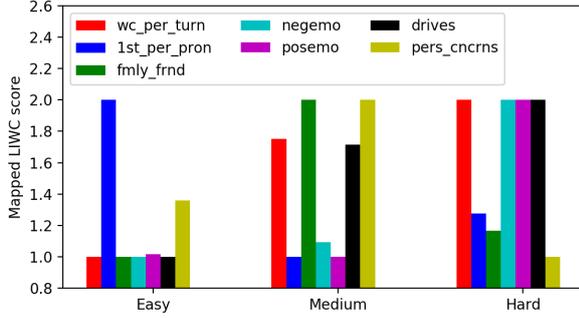


Figure 2: SD cues for three topic classes, mapped to interval [1,2] for better visualization

Feature	Highest score sessions
1st per. pron.	Getting to know, Yourself, Family
Fmly/frnd	Gathering, Family, Cooking
Neg. emot.	Driving, Growing older, Money
Pos. emot.	Yourself, Weather, Outdoors
Drives	Gathering, Life goals, Arts
Pers. concern	Growing older, Activity, Family

Table 5: Topics with the highest LIWC category scores

We therefore evaluated the average absolute sentiment value across all users for different topic classes. The results can be seen in table 3.

The results show that although the avatar’s tone remains almost the same for all classes, users tend to use stronger tones when they talk about ‘medium’ and ‘hard’ topics compared to ‘easy’ ones.

4.3 Self-disclosure

Under this heading we focus on sessions mainly concerning user’s lives, beliefs, interests, etc., expected to elicit some degree of self-disclosure. The goal is to gain insight into the dependence of self-disclosure on different topics. As mentioned earlier, there is no well-defined set of cues for measuring self-disclosure, but various studies have suggested some potentially significant ones (recall section 2). We instantiated these as follows, relying on LIWC features (Pennebaker et al.): 1)

word count per turn, 2) first person pronoun, 3) family and friends, 4) negative emotions (anxiety, anger and sadness), 5) positive emotions, 6) drives (affiliation, achievement, power, reward, risk), 7) personal concerns (work, leisure, home, money, religion and death).

We first report the LIWC-based scores of the above features in the three topic classes in table 4. To make the comparison more vivid, we linearly map the scores to [1,2] for each category independently and plot a bar graph. It can be seen that the “hard” topics contain more words per turn, and more negative and positive emotions and drives. On the other hand, people use personal pronouns more often in easy topics such as when they introduce themselves or talk about their activities. Conversation about family, friends, and personal concerns, though somewhat intimate, need not involve high self disclosure.

We also make a list of topics with the highest LIWC category scores. As can be observed in table 5, participants used the most first-person pronouns in the initial greeting session and in talking about themselves and their families. Family and friend words not surprisingly were used in “Family” and “Gathering” sessions but also when the topic was on “Cooking”. “Growing older” is among the topics where people use the most negative emotion and personal concern words.

5 Conclusion

We presented some results concerning the dialogue behavior and inferred sentiment of a group of older adults interacting with a computer-based avatar on a wide range of topics. The naturalness of the interactions, generally attested by the users, (Razavi et al., 2019) indicates that our results are meaningful. We observed that people tend to talk more when the topics are more intimate, such as life goals and the challenges of getting older, where they also use stronger emotion words—both positive and negative. Furthermore, the average response length increases as people progress along the series of interactions. These results support the use of dialogue agents with older adults in the context of difficult conversation topics. Our participants were more engaged with the agent when the conversation topics were more emotionally intense and intimate. Given the importance of effective communication during challenging conversations in later life—driving cessation, health-care, and end-of-life decision-making—our find-

ings suggest that dialogue agents could provide valuable practice and coaching to help older adults successfully navigate these challenging conversations and thereby improve both health and quality of life.

Larger studies, and branching out to other age and culture groups, will be needed to gain a fuller understanding of user behavior in such settings, and to make inferences going beyond correlations to causal analyses.

References

- Hojjat Abdollahi, Ali Mollahosseini, Josh T Lane, and Mohammad H Mahoor. 2017. A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 541–546. IEEE.
- Patricia A Arean, Michael G Perri, Arthur M Nezu, Rebecca L Schein, Frima Christopher, and Thomas X Joseph. 1993. Comparative effectiveness of social problem-solving therapy and reminiscence therapy as treatments for depression in older adults. *Journal of consulting and clinical psychology*, 61(6):1003.
- JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.
- Laura Dellazizzo, Olivier Percie du Sert, Kingsada Phraxayavong, Stéphane Potvin, Kieron O'Connor, and Alexandre Dumais. 2018. Exploration of the dialogue components in a vatar t herapy for schizophrenia patients with refractory auditory hallucinations: A content analysis. *Clinical psychology & psychotherapy*, 25(6):878–885.
- David J Houghton and Adam N Joinson. 2012. Linguistic markers of secrets and sensitive self-disclosure in twitter. In *2012 45th Hawaii International Conference on System Sciences*, pages 3480–3489. IEEE.
- Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):19.
- Juan Martínez-Miranda. 2017. Embodied conversational agents for the detection and prevention of suicidal behaviour: current applications and open challenges. *Journal of medical systems*, 41(9):135.
- Juliana Miehle, Ilker Bagci, Wolfgang Minker, and Stefan Ultes. 2019. A social companion and conversational partner for the elderly. In *Advanced Social Interaction with Agents*, pages 103–109. Springer.
- Risako Ono, Yuki Nishizeki, and Masahiro Araki. Virtual dialogue agent for supporting a healthy lifestyle of the elderly.
- 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*, pages 1241–1249. International Foundation for Autonomous Agents and Multiagent Systems.
- James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015.
- Abhilasha Ravichander and Alan W Black. 2018. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263.
- S Zahra Razavi, Lenhart K Schubert, Benjamin Kane, Mohammad Rafayet Ali, Kimberly A Van Orden, and Tianyi Ma. 2019. Dialogue design and management for multi-session casual conversation with older adults. In *IUI Workshops*.
- Alex Rinaldi, Omar Oseguera, Joann Tuazon, and Albert C Cruz. 2017. End-to-end dialogue with sentiment analysis features. In *International Conference on Human-Computer Interaction*, pages 480–487. Springer.
- Tsuyoki Ujiro, Hiroki Tanaka, Hiroyoshi Adachi, Hiroaki Kazui, Manabu Ikeda, Takashi Kudo, and Satoshi Nakamura. 2018. Detection of dementia from responses to atypical questions asked by embodied conversational agents. *Proc. Interspeech 2018*, pages 1691–1695.
- Dina Utami, Timothy Bickmore, Asimina Nikolopoulou, and Michael Paasche-Orlow. 2017. Talk about death: End of life planning with a virtual agent. In *International Conference on Intelligent Virtual Agents*, pages 441–450. Springer.