People hesitate more, talk less to virtual interviewers than to human interviewers

Lauren Faust*[*] and Ron Artstein
USC Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista CA 90094-2536, USA
Lauren.Faust@student.csulb.edu , artstein@ict.usc.edu

Abstract

In a series of screening interviews for psychological distress, conducted separately by a human interviewer and by an animated virtual character controlled by a human, participants talked substantially less and produced twice as many filled pauses when talking to the virtual character. This contrasts with earlier findings, where people were *less* disfluent when talking to a computer dialogue system. The results suggest that the characteristics of computer-directed speech vary depending on the type of dialogue system used.

1 Introduction

As computer dialogue systems become more commonplace, it becomes more relevant to ask how people's interaction with dialogue systems differs from interaction with other people. The answer, of course, will vary with different dialogue systems. This paper presents a study of comparable interview dialogues, where the interviewer is either a real person or an animated computer character controlled by a person ("Wizard of Oz" setting). Unlike earlier studies which showed that people hesitate less when talking to a computer, the present study shows that people hesitate more and produce twice as many filled pauses when talking to an animated conversational interviewer.

Existing studies show that qualitatively, human-computer dialogue exhibits many similarities to human-human dialogue. For example, an analysis of interactions between visitors and Max, an animated computer agent responding to typed input at the Heinz Nixdorf MuseumsForum in Paderborn, Germany in 2004, showed that visitors treated Max conversationally as a person, evidenced by conventional strategies of beginning and ending conversations and general cooperativeness (Kopp et al., 2005). Children have been shown to exhibit turn-taking behavior when interacting with a virtual peer (Sam the CastleMate: Cassell, 2004), and match their conversational style to that of a virtual character (Cassell et al., 2009). In an extensive literature review, Branigan et al. (2010) show that people align their speech patterns with computers at multiple levels of linguistic structure; this work also shows that the extent of alignment varies depending on whether the speaker thinks they are talking to a computer or to a person (though in the experiments cited, people were talking to computers in both belief conditions).

However, there are not many quantitative studies about the differences between comparable human-human and human-computer dialogues. Several early studies measured disfluencies in computer-directed speech. Oviatt (1995) looked at disfluencies in three corpora – a corpus of simulated human-computer interactions using speech and writing to accomplish transactional tasks such as paying bills or booking a rental car, a corpus of task-oriented telephone conversations regarding conference registration and travel arrangements, and a corpus of face-to-face dialogues and monologues giving instructions on how to assemble a water pump. The disfluency rate was significantly higher when talking to a person than when talking to a computer; within the computer-directed speech, disfluencies occurred at a higher frequency when the tasks were unconstrained rather than structured. Oviatt (1996) found that in a multimodal (speech + pen) map interaction task, disfluency rates were similar to those found in the computer-oriented speech from the previous study. Shriberg (1996) compared the frequency of disfluencies in three different corpora – a corpus of simulated human-computer interactions of air-travel planning, a corpus of real dialogues between travelers and travel agents, and the Switch-

[*] Now at California State University Long Beach

Interviewer Let's see, is there someone in your life that's been a really positive influence for you?
Participant Uh yeah, my husband, yeah.
Interviewer Yeah.
Interviewer What kind of values did you take away from him?
Participant Uh he's always uh thinking ahead and looks at the big picture and doesn't uh mull over trivial things so that's something that helped me.
Interviewer Mhm yeah, those are good traits to have.
Participant Yeah, yes.
Interviewer Um how did you guys meet?

Ellie Who's someone that's been a positive influence in your life?
Participant Uh my father.
Ellie Can you tell me about that?
Participant Yeah, he is a uh
Participant He's a very he's a man of few words
Participant And uh he's very calm
Participant Slow to anger
Participant And um very warm very loving man
Participant Responsible
Participant And uh he's a gentleman has a great sense of style and he's a great cook.
Ellie Uh huh
Ellie What are you most proud of in your life?

Figure 1: Example dialogues: face-to-face (left) and Wizard-of-Oz (right)

board corpus of general-domain telephone conversations. Here too, individuals were significantly more disfluent when talking to a person than when talking to a computer, producing more repetitions, deletions, and filled pauses. All the above studies found that the rate of disfluencies increased as the utterance length increased.

Though the aforementioned studies examine interactions between humans and computers, these dialogues cannot be said to mirror a face-to-face conversation. The computer partners are disembodied, communicating only with voice or with voice augmented by a graphical interface, and the dialogues are task-oriented rather than conversational. It is therefore not possible to draw inferences from these studies on how people will talk to conversational, embodied computer dialogue systems. Later studies used systems with more conversational characteristics: In Oviatt (2000), children aged 6–10 asked questions of computer images of sea animals, with rudimentary animations (blinking eyes) and synthesized speech; here too, the children were less disfluent when talking to the computer characters than when playing a 20-question game with an adult. A more realistic conversational agent was used in Black et al. (2009), where children aged 4–7 talked to an animated agent which used a combination of recorded and synthesized speech (Yildirim and Narayanan, 2009). In this study, children talking to the character exhibited disfluencies in fewer turns than when talking to an adult, though the effect was smaller than in the previous studies cited.

Other than Black et al. (2009) we have not found studies of comparable corpora of human-human and human-computer interaction with embodied conversational agents. The absence of such corpora is somewhat surprising, given that it has been known for several decades that people talk differently to computers and humans (e.g. Jönsson and Dahlbäck, 1988), and since human role-playing is often a preliminary step in developing conversational dialogue systems (e.g. Traum et al., 2008, section 4.3). The present study looks at a comparable corpus developed for such a purpose – a set of human-human interviews and character-human interviews in a Wizard-of-Oz setup, both collected for the eventual development of a fully automated conversational agent that will act as an interviewer, screening people for mental distress (see examples in Figure 1). In this corpus it turns out that the rate of filled pauses is higher when talking to a character than when talking to a person, suggesting that the previous results are not a general property of computer-directed speech, but rather specific to the type of dialogue systems used in the studies. The increase in disfluency when interviewed by an embodied conversational agent, compared to prior research showing a decrease in disfluencies when talking to disembodied agents, is consistent with the results of Sproull et al. (1996), who show that participants take longer to respond and type fewer words when interviewed by a talking face compared to a textual interview.

The remainder of the paper describes the corpus, the measures taken, and the differences found between human-human and character-human conversations. Our results show that patterns of conversation with disembodied, task-oriented dialogue systems do not carry over to embodied conversational agents. More generally, it is not appropriate to talk about how people talk to computers in general, because the way people talk varies with the type of dialogue system they talk to.

2 Method

2.1 Materials

We used a corpus of interviews, designed to simulate screening interviews for psychological distress, collected as part of an effort to create a virtual interviewer character. The interviews are of two types (see examples in Figure 1).

Face-to-face interviews, where a participant talks to a human interviewer situated in the same room; these are a subset of the interviews analyzed by Scherer et al. (2013) for nonverbal indicators of psychological distress.

Wizard-of-Oz interviews, where a participant talks to an animated virtual interviewer controlled by two human operators sitting in an adjacent room; a subset of these interviews were analyzed in DeVault et al. (2013) for verbal indicators of psychological distress.

The face-to-face interviews were collected during the summer of 2012. Participants were interviewed at two sites: at the USC Institute for Creative Technologies in Los Angeles, California, and at a US Vets site in the Los Angeles area. Participants interviewed at ICT were recruited through online ads posted on Craigslist.org; those interviewed at the US Vets site were recruited on-site, and were mostly veterans of the United States armed forces. After completing a set of questionnaires alone on a computer, participants sat in front of the interviewer for the duration of the interview (Figure 2); only the participant and interviewer were in the room. Interviews were semi-structured, starting with neutral questions designed to build rapport and make the participant comfortable, progressing to more specific questions about symptoms and events related to depression and PTSD (Post-Traumatic Stress Disorder), and ending with neutral questions intended to reduce any distress incurred during the interview. Participant and interviewer were recorded with separate video cameras, depth sensors (Microsoft Kinect), and lapel microphones. For additional details on the collection procedure, see Scherer et al. (2013).

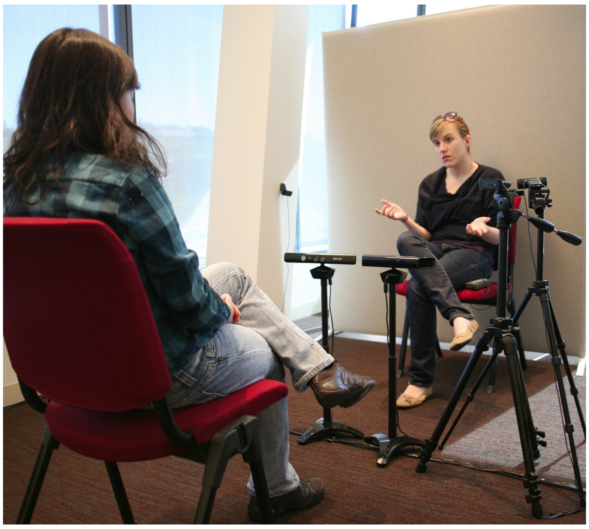


Figure 2: Face-to-face interview setup.

The Wizard-of-Oz interviews were collected in three rounds during the fall and winter of 2012–2013. All the participants were recruited through online ads posted on Craigslist and interviewed at the USC Institute for Creative Technologies. As with the face-to-face interviews, participants first completed a set of questionnaires on a computer, and then sat in front of a computer screen for an interview with the animated character, Ellie (Figure 3). No person other than the participant was in the room. The interviewer's behavior was controlled by two wizards, one responsible for the non-verbal behaviors such as head-nods and smiles, and the other responsible for verbal utterances (the two wizards were the same people who served as interviewers in the face-to-face data collection). The character had a fixed set of verbal utterances, pre-recorded by an amateur actress (the wizard controlling verbal behavior). The Wizard-of-Oz interviews were semi-structured, following a progression similar to the face-to-face interviews. Participants were recorded with a video camera, Microsoft Kinect, and a high-quality noise-canceling headset micro-

Figure 3: Ellie, the virtual interviewer.

phone.

There were small differences in protocol between the three rounds of the Wizard-of-Oz data collection. In the first round, the introductory explanation given to the participants did not explicitly clarify whether the interviewer character was automated or controlled by a person; in the subsequent rounds, each participant was randomly assigned to one of two framing conditions, presenting the character as either an autonomous computer system or a system controlled by a person. We did not find differences between the framing conditions on the measures described below, so the results reported in this paper do not look at the framing condition variable. An additional difference between the three Wizard-of-Oz collection rounds was the interview protocol, which became stricter and more structured with each successive round. Finally, with each round the character received a few additional utterances and nonverbal behaviors.

In both the face-to-face and Wizard-of-Oz conditions, each participant completed a series of questionnaires prior to the interview; these included the PTSD Checklist – Civilian Version (PCL-C) (Blanchard et al., 1996) and the Patient Health Questionnaire, depression module (PHQ-9) (Kroenke and Spitzer, 2002). There are strong correlations between the results of the two questionnaires (Scherer et al., 2013, Figure 1), so for the purpose of the analysis in this paper, we collapse these into a single assessment of distress: participants who scored positive on either of the questionnaires are considered distressed, while those who scored negative on both are considered non-distressed. In the face-to-face condition, interviewers received the results of the questionnaires prior to the interview, whereas in the Wizard-of-Oz condition, wizards were blind to the participant's distress condition.

Condition	Distressed	Non-distressed
Face-to-face	34	40
Wizard-of-Oz	59	124

Table 1: Participants and conditions.

Overall, our analysis considers the gross division of the participant population into two interview conditions (face-to-face and Wizard-of-Oz) and two distress conditions (distressed and non-distressed); see Table 1. We do not consider differences between the Wizard-of-Oz collection rounds or framing conditions, nor differences between the veteran and non-veteran populations or the individual interviewers in the face-to-face interviews. While it is known that demographic factors affect language behavior, and in particular disfluency rates (Bortfeld et al., 2001), differences between the US Vets and general population turned out non-significant on all the measures reported below, with the exception of rate of plural pronouns which was marginally significant at $p = 0.03$. Splitting the participant population into smaller groups would make it more difficult to detect the trends in the broad categories.

All the dialogues were segmented and transcribed using the ELAN tool from the Max Planck Institute for Psycholinguistics (Brugman and Russel, 2004),[1] and each transcription was reviewed for accuracy by a senior transcriber. Utterances were defined as continuous speech segments surrounded by at least 300 milliseconds of silence. For the face-to-face dialogues, both participant and interviewer were transcribed; for the Wizard-of-Oz dialogues only the participant was transcribed manually, while the interviewer utterances were recovered from the system logs.

2.2 Procedure

Several measures were extracted from the transcriptions of the interviews using custom Perl scripts.

Quantity measures: Total time of participant

[1] http://tla.mpi.nl/tools/tla-tools/elan

speech; total number of participant words; speaking rate; utterance length.

Disfluency measures: Filled pauses (*uh, um, mm*) per thousand words; percentage of utterances beginning with a filled pause.

Lexical items: First person singular (*I, me, our*) and plural (*we, us, our*) pronouns; definite (*the*) and indefinite (*a, an*) articles.

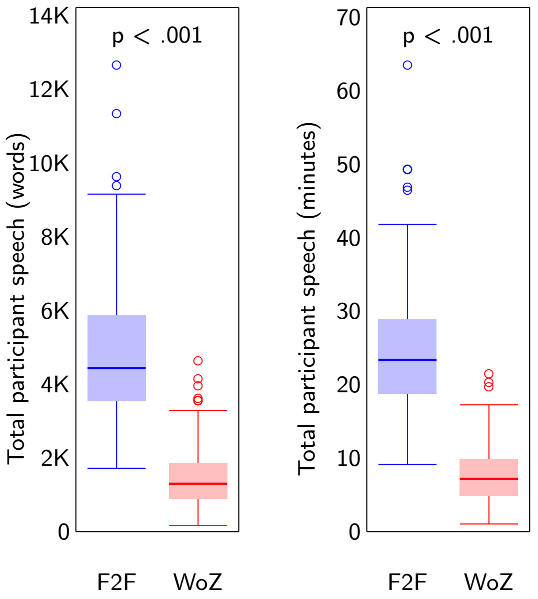
The above measures were calculated individually for each participant; we then compared the measures according to the 2×2 setup (interview condition and distress condition) described above for Table 1. Most of the significant effects we found are main effects of interview condition. Since the values typically do not follow a normal distribution, we report these effects using Wilcoxon rank-sum tests.

3 Results

3.1 Speech quantity

The face-to-face dialogues were substantially longer than Wizard-of-Oz dialogues (Figure 4): the median face-to-face dialogue participant uttered 4432 words and spoke for 23 minutes, while the median Wizard-of-Oz dialogue participant uttered only 1297 words and spoke for only 7 minutes; the differences are highly significant ($W \approx 450$, $n_1 = 74$, $n_2 = 183$, $p < .001$). The difference in speech quantity is likely due to several limitations of the wizard system. With a fixed set of utterances, the wizard runs out of things to say at some point, whereas human interviewers can engage the participants for much longer. Additionally, the human interviewer can tailor the questions to the participant's previous response, going deeper into each discussion topic than is possible for a wizard.

Not only did participants talk more in the face-to-face condition, they also used longer utterances. We calculated the mean number of words per utterance for each speaker (Figure 5, left panel): the median is 16 words per utterance in the face-to-face dialogues and 8 in the Wizard-of-Oz dialogues ($W = 1398$, $p < .001$). One possible reason for the difference is that speakers may be aligning their utterances to match the length of the interviewer's utterance. Another possible reason is the near-absence of verbal backchannels



Boxplot legend

Heavy line	Median
Shaded area	First and third quartiles
Whiskers	Data within 1.5 times interquartile range
Circles	Outliers beyond 1.5 times IQR
P-value	Wilcoxon rank-sum test

Figure 4: Speech quantity.

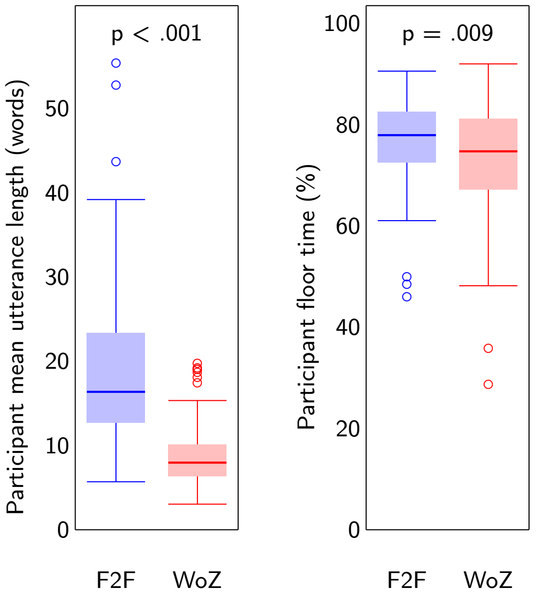


Figure 5: Utterance length and floor time.

in the Wizard-of-Oz dialogues. The wizard system did have verbal backchannels built in, but it was discovered during preliminary testing that participants tended to interpret these as an attempt by the interviewer to take the floor, and would subsequently stop speaking. As a consequence, the wizards did not use verbal backchannels during the main data collection, but only non-verbal backchannels. The verbal backchannels given by human interviewers in the face-to-face condition, in particular their ability to give specific feedback (Bavelas et al., 2000), may be a contributing factor which encourages longer participant utterances.

Participants also held the floor longer in the face-to-face condition, calculated as the proportion of participant speech duration out of total speech duration (Figure 5, right panel): median 77% of the total talking time, as compared to 75% in the Wizard-of-Oz condition; while the difference is not large, it is statistically significant ($W = 5261$, $p = 0.009$).[2]

There were also differences in speech quantity between distressed and non-distressed individuals, but only in the face-to-face condition (interaction between interview and distress conditions in a 2×2 ANOVA: $F(1,253) = 20$ for participant words, $F(1,253) = 15$ for participant time, $p < .001$ for both measures). The reason for this difference is the interview protocol: in the face-to-face dialogues, interviewers knew the participants' distress condition prior to the interview, and the protocol for interviewing distressed participants included more questions than for non-distressed participants. In the Wizard-of-Oz condition, wizards did not have access to the participants' medical condition, so the protocol was the same and there were no ensuing differences in dialogue length.

3.2 Filled pauses

Individuals in the Wizard-of-Oz condition produced filled pauses (*uh, um, mm*) at a rate almost twice that of individuals in the face-to-face condition: median 46 per thousand words in the wizard condition, 26 in the face-to-face condition ($W = 2556$, $p < .001$, Figure 6). The rate of utterances beginning with a filled pause was also significantly greater in the Wizard-of-Oz condition (median 19%) than in the face-to-face condition (median 9%; $W = 2580$, $p < .001$), possibly indicating that participants hesitated more when responding to the virtual interviewer. These findings are opposite to what is described in the literature, where people produce fewer disfluencies when talking to a voice-only, task-oriented dialogue system (Oviatt, 1995; Shriberg, 1996).

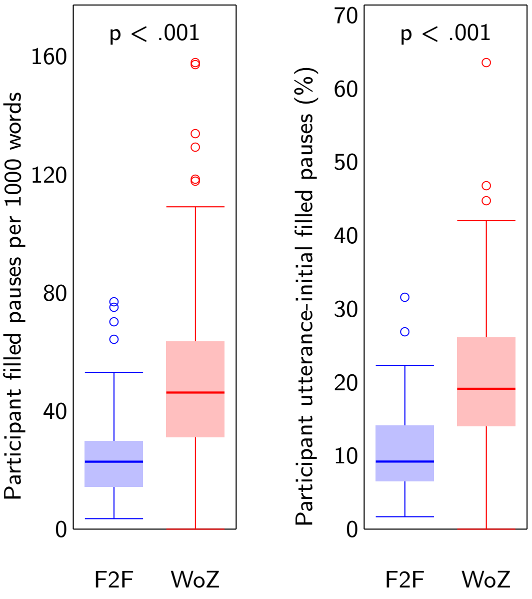


Figure 6: Filled pauses.

Even more striking is the relation between filled pause rate and utterance length. While previous literature has reported that longer utterances have higher rates of disfluency (Oviatt, 1995; Oviatt, 1996; Oviatt, 2000), our dialogues show the opposite: longer utterances have lower rates of filled pauses (Figure 7). The drop rather dramatic, starting with the one-word utterances – 38% of these in the Wizard-of-Oz dialogues and 19% in the face-to-face dialogues consist of just a filled pause. The difference between the observed pattern and the one noted in previous literature is a further indication that the current dialogues are of a different nature than the ones investigated in the prior work.

We did not find a significant difference between the filled pause rates of distressed and non-distressed individuals. Working on a portion of the same data (43 dialogues from the second round of Wizard-of-Oz testing), DeVault et al. (2013) did find a significant difference, whereby distressed individuals produced fewer filled pauses per utterance than non-distressed individuals. This discrepancy is due to the fact that the current study uses more data, and employs a different depen-

[2] We excluded 3 Wizard-of-Oz dialogues from this test because errors in logging precluded the calculation of character speech time.

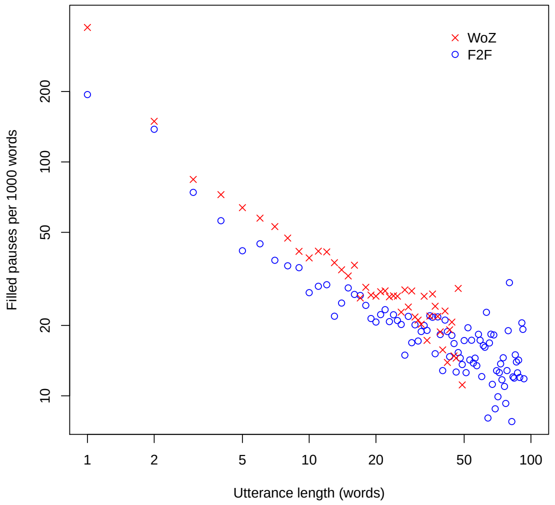


Figure 7: Relation between utterance length and filled pause rate. Data are plotted only when the corpus contains at least 10 utterances of a specified length.

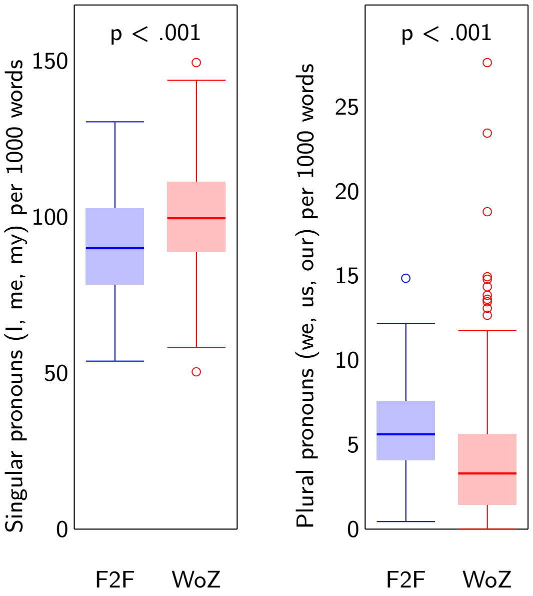


Figure 8: First-person pronouns

dent measure of disfluency (filled pauses per 1000 words rather than filled pauses per utterance). Measuring filled pauses per utterance on the full set of Wizard-of-Oz data failed to find a significant difference between distressed and non-distressed individuals, nor was a significant difference found when measuring filled pauses per 1000 words on the 43-dialogue subset.

3.3 Lexical items

An increased use of first-person singular pronouns has been linked to psychological distress in studies that compared the writing of suicidal and non-suicidal poets (Stirman and Pennebaker, 2001) and reflective essays by students (Rude et al., 2004); we tested these variables in order to see if these results carry over to dialogue. We did not find differences between distressed and non-distressed individuals or interactions between interview condition and distress condition, but we did find differences between the face-to-face and Wizard-of-Oz dialogues: first person singular pronouns (*I, me, my*) were used at a higher rate in the Wizard-of-Oz condition (median 100 per thousand words compared to 90 in the face-to-face condition, $W = 4608$, $p < .001$), whereas first-person plural pronouns (*we, us, our*) were used at a higher rate in the face-to-face condition (median 6 per thousand words compared to 3 in the Wizard-of-Oz condition, $W = 4075$, $p < .001$; see Figure 8). We do not have an explanation for these differences, and we cannot say whether they reflect a general difference between human-human and human-character interactions, or if they are caused by specific properties of the experimental setup in the two conditions. A sampling of the plural pronouns showed that they are primarily exclusive, that is they refer to the speaker and someone else but not the interviewer.

Face-to-face And uh I hooked up with uh somebody who runs this company uh at a party and uh we started talking and uh he offered me the job.

Wizard-of-Oz I have a stepfather and a half-brother we get along okay but we're not very close.

Other uses of *we* were generic.

Face-to-face And I'm a passionate believer in our trying to get our country going straight. I think we're we're going the wrong way and I don't know there's any way to stop it.

Wizard-of-Oz That's one of the things that took me a number of years to master though were my relaxation skills, I think that's a key thing and I think as we mature, as we learn how to do that, I wish I'd learned how to do that.

However, at least one participant referred to the virtual interviewer Ellie with an inclusive *we*:

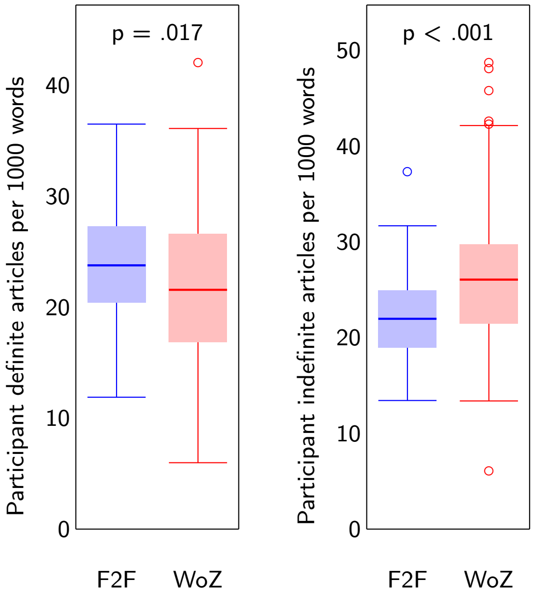


Figure 9: Articles.

Wizard-of-Oz Well, in the last few minutes since we started talking about depressing stuff, I starting to feel a little more down.

We also found a difference in the use of articles between the face-to-face and Wizard-of-Oz conditions: face-to-face dialogues contained more definite articles than Wizard-of-Oz dialogues (medians 24 and 22 per thousand words, $W = 5481$, $p = .02$), whereas the opposite is true for indefinite articles (medians 22 and 26 per thousand words, $W = 4263$, $p < .001$; Figure 9).

4 Discussion

Two main findings emerge from the present study. One is that human interviewers are able to engage participants in much longer conversations than Wizard-of-Oz characters. This is not surprising, given that the animated character has only a fixed number of utterances. Even in the short dialogue samples in Figure 1 above we can see how the human interviewer can tailor follow-up utterances to the participant's contributions, while the wizard-controlled character can only use generic follow-ups and has to move on when these are exhausted.

The second finding is that participants produce more filled pauses when talking to the animated interviewer than when talking to a human interviewer. This finding is important because it is the opposite of earlier results about computer-directed speech. Of course, the earlier results are from a very different kind of dialogue system – a disembodied, task-oriented dialogue interface as opposed to an animated conversational character. Nevertheless, these results have been taken to apply to computer-directed speech in general (e.g. Corley and Stewart, 2008, page 591: "Speakers tend to be more disfluent overall when addressing other humans than when addressing machines," making reference to Oviatt, 1995). The present study shows that the results from disembodied task-oriented systems do not carry over to conversational dialogue systems, and more generally that computer-directed speech is not a unitary phenomenon, but that it varies depending on the computer system that the speech is directed to.

As mentioned in section 2.1, the face-to-face and Wizard-of-Oz dialogues were collected with the eventual goal of creating a fully automated character capable of interviewing people about mental distress. Experiments with an automated prototype are currently underway, and we hope to have access to dialogues between people and a fully automated character soon. Having a corpus with three types of comparable interview dialogue – human, human-controlled, and automated interviewers – will hopefully shed additional light on the question of the characteristics of computer-directed speech.

Acknowledgments

The first author was supported by the National Science Foundation under grant 1263386, "REU Site: Research in Interactive Virtual Experiences" (PI: Evan Suma).

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract N66001-11-C-4006, DARPA contract W911NF-04-D-0005, and the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, December.

Matthew Black, Jeannette Chang, Jonathan Chang, and Shrikanth S. Narayanan. 2009. Comparison of child-human and child-computer interactions based on manual annotations. In *Proceedings of the Workshop on Child, Computer and Interaction*, pages 2065–2068, Cambridge, MA, November.

Edward B. Blanchard, Jacqueline Jones-Alexander, Todd C. Buckley, and Catherine A. Forneris. 1996. Psychometric properties of the PTSD checklist (PCL). *Behaviour Research and Therapy*, 34(8):669–673, August.

Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147, June.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2065–2068, Lisbon, Portugal, May.

Justine Cassell, Kathleen Geraghty, Berto Gonzalez, and John Borland. 2009. Modeling culturally authentic style shifting with virtual peers. In *ICMI-MLMI '09: Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 135–142, Cambridge, Massachusetts, November. ACM.

Justine Cassell. 2004. Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Language and Linguistics Compass*, 2(4):589–602.

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, Metz, France, August. Association for Computational Linguistics.

Arne Jönsson and Nils Dahlbäck. 1988. Talking to a computer is not like talking to your best friend. In Thore Danielsen, editor, *Proceedings of The First Scandinivian Conference on Artificial Intelligence*, pages 53–68, Tromsø, Norway, March.

Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide – design and evaluation of a real-world application. In Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist, editors, *Intelligent Virtual Agents: 5th International Conference, IVA 2005, Kos, Greece, September 12–14, 2005 Proceedings*, volume 3661 of *Lecture Notes in Artificial Intelligence*, pages 329–343, Heidelberg, September. Springer.

Kurt Kroenke and Robert L. Spitzer. 2002. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9):509–515, September.

Sharon Oviatt. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, 9(1):19–35.

Sharon Oviatt. 1996. Multimodal interfaces for dynamic interactive maps. In *Conference on Human Factors in Computing Systems: Common Ground (CHI '96)*, pages 95–102, Vancouver, BC, Canada, April. ACM.

Sharon Oviatt. 2000. Talking to thimble jellies: Children's conversational speech with animated characters. In *Sixth International Conference on Spoken Language Processing, (ICSLP 2000)*, volume 3, pages 877–880, Beijing, China, October. ISCA.

Stephanie S. Rude, Eva-Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *10th IEEE International Conference on Automatic Face and Gesture Recognition*, Shanghai, China, April.

Elizabeth Shriberg. 1996. Disfluencies in Switchboard. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of ICSLP 96, the Fourth International Conference on Spoken Language Processing*, volume Addendum, pages 11–14, Philadelphia, October.

Lee Sproull, Mani Subramani, Sara Kiesler, Janet H. Walker, and Keith Waters. 1996. When the interface is a face. *Human-Computer Interaction*, 11(2):97–124, June.

Shannon Wiltsey Stirman and James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522.

David Traum, William Swartout, Jonathan Gratch, and Stacy Marsella. 2008. A virtual human dialogue model for non-team interaction. In Laila Dybkjær and Wolfgang Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, chapter 3, pages 45–67. Springer, Dordrecht.

Serdar Yildirim and Shrikanth Narayanan. 2009. Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):2–12, January.