Common Ground in Dialogue with a Gendered Humanoid Robot*

Aaron Powers, Adam Kramer, Shirlene Lim, Jean Kuo,
Sau-lai Lee, Sara Kiesler

Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes, Pittsburgh, PA 15232, USA

apowers@cmu.edu, adk@andrew.cmu.edu, sllim@andrew.cmu.edu,
cjkuo@andrew.cmu.edu, slleeh@hkusua.hku.hk, kiesler@cs.cmu.edu

***Abstract* – Research on human-human interaction shows that people estimate each others' knowledge to find where they have a common ground of understanding. They adapt their speech to this common ground. In particular, people explain themselves more to others when they have less common ground. We demonstrate that the same process can occur in human-robot interaction. In discussing dating, women participants explained dating norms for a woman more to a "male" robot than to a "female" robot, and men generally showed the opposite behavior. We discuss the development of the dialogue that produced common ground, and what our results imply for human-robot interfaces and interaction design.**

Index Terms* – *human-robot interaction, social robots, humanoids, communication, dialogue

INTRODUCTION

To communicate effectively with other people, we need to have a reasonably accurate idea about what specific other people know [1]. In this paper, we argue that some of the processes by which people get to know others and communicate with them effectively may also apply to human-robot interaction.

An obvious starting point for people to build a model of what a robot knows is what they themselves know, or think they know. Their own knowledge acts as a default or "anchor" for estimating the knowledge of others [2]. People can also use social context cues to build a model of what a robot knows. Social cues point to social categories (e.g., gender, age, group membership, and so forth), which in turn help people estimate others' knowledge [3]. For example, if a robot is a humanoid, people might imagine that it shares some characteristics of humans, for example, a tendency to prefer friends to strangers. Or, if the robot is known to originate in New York or Hong Kong, people might conclude that it has knowledge about these localities [4]. People's mental models of a robot are important because these models establish the amount and nature of common ground between them and the robot. This common ground, in turn, is likely to affect how people communicate with the robot and what they expect in return. Speakers design messages to be appropriate to what they assume to be the knowledge of the recipients [5]. For instance, people represent information more elaborately if they have to communicate it to others who know nothing about the subject matter [6, 7].

In this paper, we argue that a gendered robot, that is, a robot that emits gender cues, will elicit a mental model of the robot such that people estimated it to have knowledge usually specific to men or women. For example, a robot that speaks with a feminine (high frequency) voice or that looks feminine might be estimated to have knowledge of women's clothing sizes and women's sports celebrities. A robot that speaks with a masculine (low frequency) voice might be estimated to have knowledge of men's clothing sizes and men's sports celebrities.

If group membership or social category is a cue to what people know [1], we might expect that people will estimate a female robot to have more knowledge of dating practices than a male robot will. They could derive this estimate from the general case that women are more knowledgeable about social practices and have more social skill than men do [8]. If so, when people describe a dating norm to a robot, they should assume they will have to explain the norm less to the female robot than to the male robot, because the female robot already shares some of this knowledge.

We further argue that if the people interacting with the gendered robot are the same gender as the robot seems to be, then they will assume there is more common ground between them and the robot than if they are not the same gender as the robot. (Such an assumption would not be entirely silly. Dawes [9] showed that people do well, statistically speaking, to take their own opinions or knowledge as representative of that of the group to which they belong.) The more the common ground between the person and the robot (as when they share the same gender), the less they will need to explain subject matter that they both are assumed to know in common. Hence, in the prior example, if women assume that they and a "female" robot both understand dating norms for a woman (for example, that she is not expected to pay for a date), then when asked to articulate the norm, women will explain it with less elaboration than if the robot is "male." Likewise, men may explain dating norms for a man to a male robot less than to a female robot.

We tested these predictions because, if valid, they have significant implications for understanding and designing human-robot social interaction. For example, the theory implies that people who interact with a gendered humanoid

* This work is supported by NSF Grant #IIS-0121426.

robot do not approach the robot *tabla rosa*, but rather hold a default mental model of the robot's knowledge and that the model influences their assumed common ground with the robot and their behavior with the robot. The mental model, of course, may be highly inaccurate. For example, people may greatly overestimate their common ground with a robot in many domains. Designers can affect these models and assumptions of common ground in appropriate directions by the way they develop the robot's appearance, behavior, social context, and responses to people.

RELATED WORK

The last decade has seen a number of projects involved in the construction of social robots, that is, robots that engage in social interaction with people. Thus, Sparky [10] and Kismet [11] used facial features, facial expression, movement, and sounds to convey attention to the observer and to the observer's responses. Museum robots [12 - 14] have been designed to traverse museum spaces, speak out loud, convey commands (such as "make way") and generally to provide display information and amusement for visitors. More recently, Valerie, a receptionist-robot, engages in a dialogue with people, giving them information (such as the location of people in offices) and entertaining them by telling stories [15]. Robovie [16] is a child-sized robot in Japan who speaks English with school children, recognizes them, and engages them in one-to-one games. The Nursebot robot, Pearl, the same robot we used in the study described in this paper, was initially developed to interact with older people who may need help to remain independent in their homes [17]. None of this work directly examines people's mental models for a robot, in particular, their mental model of what the robot knows, and how the model affects their interactions with the robot. Our purpose is to demonstrate this influence and to show how a humanoid robot's design as female or male affects the human-robot interaction.

METHOD

We tested the theory and predictions in an experiment in which young adults of both genders engaged in a one on one dialogue with a humanoid robot. Because without any interaction, we could not test differences in interaction, we had first to develop an interaction design and dialogue for a robot that novice participants would find engrossing and involving. Another requirement was to insure that participants understood the robot's questions and responses, and that the robot understood the participants' responses. A third requirement was the topic would be one in which participants had highly predictable knowledge and established opinions. We chose the topic of "first dates" because almost all young adults have personal knowledge of dating practices and because there are well-established schemas for behavior of women and men on first dates [18]. Indeed, norms for first dates have changed little since the 1950s [19].

The experimental setup was one in which the participant engaged in a dialogue with a robot who was presented as either female (feminine voice, red lips) or male (male voice, uncolored lips). The male or female robot told the participant that it was training to be a dating counselor, and that it needed advice about what typically happens on dates. It then asked various questions about events that transpire on a first date, and it responded to what the participant typed. The dialogue was scripted to begin with general questions about dating, such as where people meet others and about the appropriateness of behaviors such as dating a boss or coworker. As the dialogue progressed, the robot introduced a hypothetical couple, "Jill" and "John" who were about to go on a first date. The robot asked the participants a series of questions about each, such as whether John should call Jill back if she was busy the first time he called, or if Jill should bring John flowers. These questions about Jill and John are the focus of this article because they illuminate how women and men participants talked with a male or female robot differently depending on whether they were talking about a woman (Jill) or a man (John).

A. Experimental Design

The experimental design was a 2 X 2 X 2 factorial with two between groups factors and one within groups factor. The first two factors were participant gender and robot gender. The third factor was a within subjects factor where all participants answered some questions about Jill and John.

B. Participants

Thirty-three native American-English speakers from Carnegie Mellon participated for US$10 cash as payment (17 males, 16 females; average age 21 years).

C. Procedure

When participants arrived at the experimental lab, an experimenter told them he/she was creating a dating service for Carnegie Mellon students, and that participants' conversation would help train the robot's AI system to give people better advice. Approximately half of the sessions were run by a male experimenter and half by a female experimenter. (We analyzed our statistically to determine the effect of the gender of the experimenter, and found the experimenter's gender had no effect on the results.)

Participants conversed with the robot through an interface like that of Instant Messaging. The IM interface was on the screen on the robot's chest, with a keyboard on which the participants typed. The robot used Cepstral's Theta for speech synthesis, and its lips moved as it spoke [20]. The text also showed on the screen, in line with the participant's responses, as in IM interfaces.

The questions the robot asked were adapted primarily from Laner & Ventrone's studies of dating norms [18, 19]. In [18, 19], Laner & Ventrone conducted two studies asking students about what events typically happen on a first date, and who does them – the man, the woman, both or either. We turned twelve of the events with the largest gender difference into a scenario about "John and Jill," two hypothetical individuals who were interested in each other. Six of the items were most commonly thought to be performed by men (e.g. "decide on plans by yourself", 61% say men, 6% say women), and six were most commonly performed by women (e.g. "buy new clothes for date," 2% say men, 75% women).

The robot asked some questions about what Jill should do and some questions about what John should do. In each case, the robot asked some of these in a way that supported a gender stereotype (e.g. "Do you think that John should make the plans for the date?"), and some in a way that was the reverse of the stereotype (e.g. "Do you think it's appropriate for John to buy new clothes for a first date?"). The questions about Jill and John were embedded in other questions about dating behavior and norms (such as the wisdom of Internet dating), so as not to create suspicion about our interest in gender in this experiment.

After chatting with the robot, participants filled out a survey with a variety of questions about themselves and the robot. Our main focus in this report is participants' ratings of the robot's gender (as a check on the manipulation) and their ratings of the robot's personality and humanlikeness. Ratings of masculinity and femininity were a subset of Bem's Sex-Role Inventory [21]. We used the personality and humanlikeness ratings to address the question of whether responses to the robot might be caused by perceptions that the robot's personality or humanlikeness differed when it was "female" or "male."

D. Dialogue

The robot interpreted and responded to the participant using a customized variant of the Alice chat-bot [22] (http://www.alicebot.org), a publicly available pattern-matching text processor. As noted above, we had to develop the dialogue so that all participants understood the robot and so the robot would be able to respond appropriately to whatever participants said. To accomplish this result, we developed and refined the dialogue, and carried out two pilot studies (one with an animation of a robot and one with the actual robot).

The first version of the dialogue created very large variability across participants both in the amount and content of their speech. Some participants did respond to the robot's questions but asked for clarifying information; others did not answer at all and had to be prompted to answer. The variability in responses was very large, due to some people asking long clarifying questions and others not answering at all. One reason for this variability across people is that, initially, the chat-bot often did not respond well to participants' questions. For example, when the robot asked "Should Angela/Anthony go to a club?" some participants asked, "Can he/she dance?" In each successive test, we tailored the chat-bot's responses towards the questions and comments that participants made, especially dropping dialogue that participants did not understand.

Most questions the robot asked required a positive or negative reply. However, most people do not say a simple "yes" or "no" (e.g., see Fig. 1). To allow the robot to understand most replies, we compiled hundreds of variants of common responses from participants' responses in the pilot tests, such as "that would be nice," and "of course not." We also added an ability to respond to question-specific answers, such as "they should split the check," to allow even more comprehension by the robot. When the participant uttered something vague like "maybe," or "only if …," or when the participant otherwise failed to answer the question in a manner the robot could understand, the robot prompted the participant, "Please rephrase that," or "Please be more specific, and "tell me whether it would be OK for them to date if John was Jill's boss."

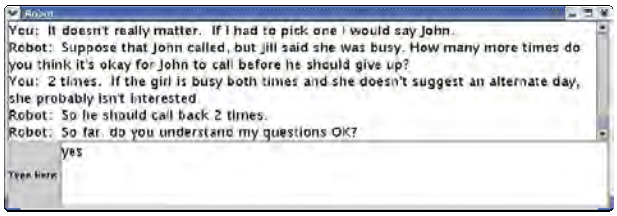

Fig. 1. IM-like chat interface, with responses from a pilot test.

Participants in the pilot studies commonly made spelling and grammatical errors, and did not correct themselves, such as saying "shoudl," or "No, is Jill and John ike each other and Jil is comfortable with asking him on a date." To fix this problem, we added the Linux Aspell spell checker to find many spelling errors and automatically correct them in the robot's interpreter. Thus when the participant spelled something wrong, the robot was still able to interpret it, and if the robot repeated the participant's words, the words were spelled correctly by the robot. As a result of these improvements in the robot's script and interpreter, the number of nonresponses by participants declined precipitously.

Although branching on the participants' responses made the interaction feel more fluid, some of the branches were boring or redundant and branches tend to complicate statistical analysis. Therefore for the main experiment we shortened the script for the robot from our original 1026 words in 65 sentences to 876 words in 50 sentences, 11 branches were reduced to only 3, and to clarify questions better the number of words in each question increased from 16.3 to 17.5.

Fig. 2 The robot talking with a participant.

E. Analyses

The dependent variable in this study was how much the participants said to the robot about what Jill's and John's behavior should be before, on, and after their first date. We used the Text Analysis and Word Counts program [23] to

count the number of words the participants used to communicate with the robot. For the text in response to each statement or question of the robot, TAWC counts the number of words in that text. A count of total words used in response to the Jill and John questions was computed by summing the total words used in response to these questions. To normalize the counts, which were skewed and left censored (a person can't say fewer than zero words to any question), logs of the totals were computed and the data were centered. The result is a standardized measure of the log of total words spoken about Jill and about John.

The data were analyzed using analysis of variance with two between factors (gender of participant and gender of robot) and one within factor (words about Jill and words about John).

RESULTS

We conducted analyses to measure how participants perceived the robot then to test the hypotheses about how much men and women participants said to the robot about Jill and John's dating behavior.

A. Participants' Perceptions

The first analysis was a check on the manipulations. That is, did the participants perceive the robot to be gendered? We asked participants a write-in question about the gender of the robot. The result was highly significant (chi square = 40, $p < .0001$). Sixteen of 17 participants in the female robot condition said the robot was female and one said "female?" In the male robot condition 14/16 participants said the robot was male, one said female, and one said "male?" Thus the female robot gave a slightly stronger impression of gender (but insignificantly so). We also asked participants to respond to a pair of 5- point rating scales (1 = low, 5 = high) asking how masculine and how feminine the robot was. The interaction of robot gender and ratings on the scale was highly significant, and there were no differences in this respect between men and women participants. Participants rated the female robot an average of 3 on the feminine scale and 2.2 on the masculine scale, and they rated the male robot an average of 2.1 on the feminine scale and 3.6 on the masculine scale ($F = [1, 29] = 25$, $p < .001$).

The next analysis was conducted to check on whether there were robot gender or participant gender differences in perceptions of the robot's speech skills. Three rating scales (1 - 5) addressed this question: the robot's speech quality, the robot's response time, and the robot's conversation skill. There were no differences due to robot or participant gender. On average, participants rated the robot's speech quality 3.3, response time 2.7, and conversation skill 2.8. These scores are lower than the ratings (approximately 3.5 – 4) that people give to other people or to themselves, but higher than in the previous version of our dialogue development (scores of 2.4 to 3.2).

We next analyzed data from several items about the robot's personality. We used a scale measuring extraversion (cheerful, attractive, happy, friendly, optimistic, warm) because extraverts tend to elicit more talk from other people. We found no differences due to robot gender. In general, the robot was seen as moderately extraverted.

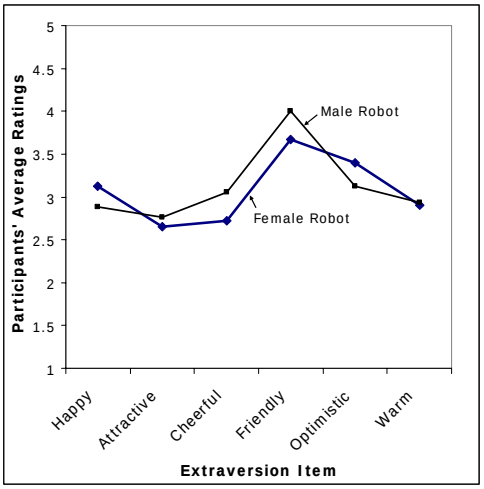

Fig. 3. Ratings of the robot's extraversion.

Other items measured the robot's dominance, compassion, and likeability. In these items, most ratings were the same across robot gender and participant gender, and in moderate ranges of the scale. However, men's ratings of the female robot were significantly lower than either women's ratings of either gendered robot or men's ratings of the male robot. Thus, men rated the female robot as lower in leadership and higher in dominance ($p < .001$), as somewhat less tender and compassionate ($p < .07$), and as marginally less likeable ($p > .10$). Because of these differences, we examined whether participants' ratings influenced how much they talked with the robot. We found that their ratings of the robot's assertiveness, compassion, and likeableness were negatively correlated with amount of talking (about $r = .20$) so we used these ratings as control variables in the subsequent analyses. Use of these control variables does not change the results.

B. Participants' Talk

As noted above, we measured the number of words that men and women participants used in communicating with the male or female robot about Jill's and John's appropriate behavior on a first date. We predicted, first, that because females are expected to know more about sociability and social norms than men, participants would estimate more common ground about dating norms with a female robot than a male robot, and would feel less need to explain dating norms to the female robot. Hence they should use fewer words too answer questions about Jill and John's first date when speaking with the female robot than when speaking with the male robot. In addition, women should estimate more common ground between them and the female robot when talking about Jill. By contrast, men should estimate more common ground between them and the robot when talking with the male robot about John.

Overall we found a significant triple interaction of participant gender, robot gender, and Jill vs. John questions ($F [1, 25] = 4$, $p = .05$). These results reflected the following:

1. Participants said fewer words to the female robot than to the male robot, as predicted (Figs. 4 and 5).

2. Women said the fewer words to the female robot than to the male robot about Jill, as predicted. They also talked the most when talking with the male robot about Jill, which fits with the theory. (That is, they are sharing knowledge with another who does not know what they know.) In this case they know a lot about what Jill should do and need to explain the norms to the male robot (Fig. 4).

3. Men said slightly fewer words to the male robot about John than they said to the female robot about John, as predicted (Fig. 5). Contrary to prediction, they said more to the female robot about Jill than John. We believe this result may have obtained if men felt less certain of their common ground with the female robot when talking about a woman's dating.

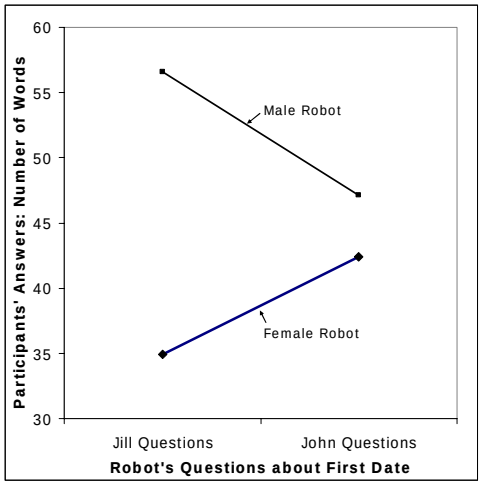

Fig. 4. Women participants' responses to the robot.

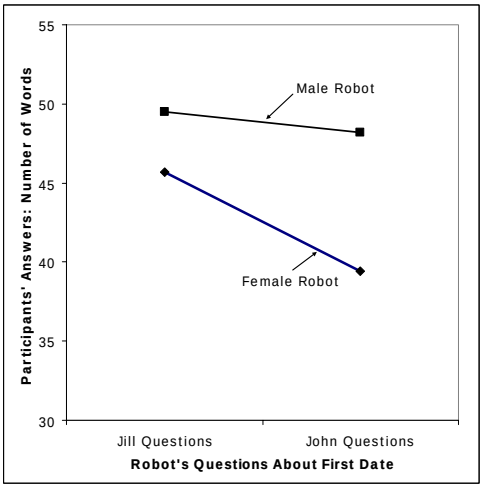

Fig. 5. Men's responses to the robot.

DISCUSSION

In summary, participants in this controlled experiment engaged in a one-on-one dialogue about (human) dating practices with an interactive humanoid robot. The ostensible purpose of this dialogue was to give the robot more knowledge about dating so it could perform as a dating counselor. Half of the participants interacted with a "female" robot with pink lips who asked them questions about dating in a feminine voice; half of the students interacted with the same robot, but it was "male"—spoke with a masculine voice and had grey lips (same as its body). We predicted that the female robot would be presumed to know about dating (more than the male robot) and hence people would explain themselves less to the female robot than to the robot. We did find this to be the case, and the result held when we did or did not control for how much participants liked the robot. We also found, among women, that they said least to the female robot about what a woman, "Jill" should do on a first date and most to a male robot about what a man, "John" should do on a first date. The results for men only partially supported the hypotheses, perhaps because men did not know as much about dating norms.

This work has at least three significant design implications for human robot interaction. First, the theory says that people will make assumptions about the knowledge of a robot based on what they themselves know and will relate to the robot in relation to what they know. Hence designers cannot assume that people approach a robot *tabla rosa* but instead with a mental model. Second, people will use the robot's outward appearance, overt behavior, and context cues to modify their default mental model of what a robot knows. Hence, designers can manipulate a robot's appearance, behavior, and context to convey the robot's knowledge or they can design a robot whose cues adapt to different user models. Third, because people will adjust their behavior to a robot depending on their common ground with it, designers will need to make decisions about whether the robot's cues appropriately or inappropriately convey common ground.

A. Future Work

Because this study represents the first demonstration of a common ground effect in human-robot interaction, we must regard it as preliminary. We believe there are many worthwhile domains to explore in seeking replication and extension of the theory to human-robot interaction, for example, whether people find common ground with a robot's emotional state, preferences, or decision biases. This work may also lead to some new ways that designers can adapt dialogue systems such that people and robots will communicate more clearly.

ACKNOWLEDGMENTS

We thank the People and Robots, Nursebot, and Social Robots project teams for their suggestions and help in designing the human-robot interactions used in this study.

REFERENCES

[1] R. S. Nickerson, "How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others," *Psychological Bulletin*, vol. 125, no. 6, pp. 737-759, 1999.

[2] A. Tversky and D. Kahneman, "Judgement under uncertainty: Heuristics and biases," *Science*, vol. 185, pp. 1124-1131, September 27, 1974.

[3] M. Ross and D. Holmberg, "Recounting he past: Gender differences in the recall of events in the history of a close relationship." In J. M. Olson and M. P. Zanna (Eds.), *Self-inferences processes*: *The Ontario Symposium*, Vol. 6, Hillsdale, NJ: Erlbaum, 1988, pp. 135-152.

[4] S-L. Lee, S. Kiesler, I.Y. Lau, and C.Y. Chiu, "Human Mental Models of Humanoid Robots," unpublished.

[5] H. H. Clark, "*Arenas of language use*," Chicago: University of Chicago Press, 1992.

[6] E. A. Issacs and H. H. Clark, “References in conversation between experts and novices,” *Journal of Experimental Psychology:General*, vol. 116, no. 1, pp. 26-37, 1987.

[7] S. Fussell and R. Krauss, “Coordination of knowledge in communication: Effects of speakers’ assumptions about what others know.” *Journal of Personality and Social Psychology*, vol. 62, pp. 378-391, 1992.

[8] W. Wood, N. Rhodes, M. Whelan, M., “Sex differences in positive well-being: A consideration of emotional style and marital status.” *Psychological Bulletin*, vol. 106, pp. 249-264, 1989.

[9] R. M. Dawes, “Statistical criteria for establishing a truly false consensus effect,” *Journal of Experimental Social Psychology*, vol. 25, pp. 1-17, 1989.

[10]M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, “Ex with Sparky: A social robot,” *Proceedings of the Workshop on Interactive Robot Entertainment*, 2000.

[11]C. Breazeal and B. Scassellati, “How to build robots that make friends and influence people,” *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Knyoju, Japan, 1999.

[12]W. Bugard et al, “Experiences with the interactive museum tour-guide robot,” *Artificial Intelligence*, vol. 114, nos. 1-2, pp. 3-55, 1999.

[13]M. Montemerlo, J. Pineau, N. Roy, S. Thrun and V. Verma, “Experiences with a mobile robotic guide for the elderly,” *18th National Conference on Artificial Intelligence*, pp. 587-592, 2002.

[14]T. Willeke and C. Kunz and I. Nourbakhsh, “The History of the Mobot Museum Robot Series: An Evolutionary Study,” *Proceedings of FLAIRS 2001*, 2001.

[15]R.Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K.Snipes, A. Shultz_, and J. Wang, “Designing robots for long-term social interaction.”

[16]T. Kanda, T. Hirano,and D. Eaton, “Interactive robos as social partners and peer tutors for children: A field trial, *Human Computer Interaction*, vol. 19, pp. 61-84, 2004.

[17]M. Montemerlo, J. Pineau, N. Roy, S. Thrun and V. Verma, "Experiences with a mobile robotic guide for the elderly", *18th National Conference on Artificial Intelligence*, pp. 587—592, 2002.

[18]M.R. Laner, and N.A. Ventrone, “Egalitarian daters/traditionalist dates,” *Journal of Family Issues*, vol. 19, no. 4, pp. 468-477, July 1998.

[19]M.R. Laner, N.A. Ventrone, “Dating scripts revisited,” *Journal of Family Issues*, vol. 21, no. 4, pp. 488-499, May 2000.

[20]K.A. Lenzo, and A.W. Black, *Theta*, Cepstral, http://www.cepstral.com

[21]S.L. Bem, *Bem Sex-Role Inventory*, Palo Alto: Consulting Psychologists Press, Inc., 1976.

[22]R. Wallace, *Alice*, ALICE Artificial Intelligence Foundation, http://www.alicebot.org/

[23]A.D.I. Kramer, S.R. Fussell, and L.D. Setlock, “Text analysis as a tool for analyzing conversation in online support groups,” *Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems*, pp. 1485-1488, April 2004.