

CROSS-CULTURAL BELIEVABILITY OF ROBOT CHARACTERS

MAXIM MAKATCHEV

CMU-RI-TR-12-24

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Robotics

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

February 2013

Thesis Committee: Reid Simmons (Chair), Michael Agar,
Justine Cassell, Illah Nourbakhsh, Candace Sidner

Copyright © 2013 by Maxim Makatchev. All rights reserved.

Maxim Makatchev: *Cross-Cultural Believability of Robot Characters*
Submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in Robotics, ©
February 2013

To you.

ABSTRACT

Believability of characters is an objective in literature, theater, animation, film, and other media. Virtual characters, believable as sharing their ethnic background with users, improve their perception of the character and, sometimes, even their task performance. Social scientists refer to this phenomenon as homophily—humans tend to associate and bond with similar others. Homophily based on ethnic similarity between humans and robots, however, has not previously been tested, in part due to the difficulties of endowing a robot with ethnicity. We tackle this task by attempting to avoid blatant labels of ethnicity such as clothing, accent, or ethnic appearance (although we control for the latter), and instead aim at evoking ethnicity via more subtle verbal and nonverbal behaviors.

Until now, when designing ethnically-specific virtual agents, their behaviors have been typically borrowed from anthropological studies and cultural models. Other approaches collect corpora of human interactions in target contexts and select maximally distinctive behaviors for further implementation on a virtual character. In this thesis, we argue that both behaviors that signal differences between an anthropologist and the target ethnicity (rich points), as well as maximally distinctive behaviors between target ethnicities, may vary on their ability to evoke ethnic attribution. We address this discrepancy by performing an additional evaluation of the candidate behaviors on their salience as ethnic cues via online crowdsourcing. The most salient ethnic cues are then implemented on the robot for a study with colocated participants.

This methodology has allowed us to design robot characters that elicit associations between the robot’s behaviors and ethnic attributions of the characters as native speakers of American English, or native speakers of Arabic speaking English as a foreign language, by members of both of these ethnic communities. Although we did not find evidence of ethnic homophily, we believe that the suggested pathway can be used to create robot characters with a higher degree of perceived similarity, and better chances of evoking homophily effect.

ACKNOWLEDGMENTS

I would like to thank my advisor, Reid Simmons, for his continual support. His patience, insight, and encouragement were crucial for my survival as a graduate student and actually enjoying the process.

I would also like to thank my thesis committee members for their time and expertise: Mike Agar, Justine Cassell, Illah Nourbakhsh, and Candace Sidner. Working with them made me feel like playing basketball with Michael Jordan (on my team).

My work has benefited from interactions with Nik Nailah Binti Abdullah, Hatem Alismail, Greg Armstrong, Eleanor Avrunin, Lucia Castellanos, Michael Chemers, Martha Clarke, Stanislav Funiak, Andrew Harris, Milos Hauskrecht, Breelyn Kane, Gunhee Kim, Heather Knight, Alex Kravets, Jun-young Kwak, Min Kyung Lee, Daniel Leeds, Diane Litman, Suzanne Lyons Muth, Matthew Marge, John McPhee, Nik Melchior, Sanako Mitsugi, Jack Mostow, Ann Mundell, Minh Hoai Nguyen, Chris Niessl, Ram Ravichandran, Antonio and Dana Roque, Alicia Sagae, Brennan Sellner, Christian Sommer, Aaron Steinfeld, Ming Sun, Eric Tschetter, Kei Usui, Jiuguang Wang, Suzanne Wertheim, and other friends and colleagues. Matthew Marge, participants of Dialogs on Dialogs seminar at CMU, and my classmates have often been the first audiences subjected to the rough treatment by early versions of my presentations.

I would like to acknowledge students, staff, and faculty of Carnegie Mellon Qatar and Education City for their support. In particular, I am grateful to Ameer Ayman Abdulsalam, Kinh Elizabeth Ai, Brett Browning, Imran Aslam Fanaswala, Swapnil Joshi, Wael Mahmoud Gazzawi, Carol Miller, Suhail Rehman, Rana Rwaished, Majd Sakr, and Micheline Ziadee.

I would not be able to grow outside my shell if not for people who gave me, as an outsider, a chance. Michele de la Reza took me in and Peter Kope bore with me at their dance classes, as my dance classmates showed great support. Mark Thompson opened to me the world of human movement. Shannon Sindelar and Dan Efros let me peek into their work on the production of Macbeth. Manuel Blum, Illah Nourbakhsh, and Steven Rudich let me sit in on their classes before I was affiliated with CMU. Dmitry Demin, Sherman Y. T. Lang, S. K. Tso, Pamela Jordan, Kurt VanLehn, Jack Mostow, Daniel Neill, and Reid Simmons all betted on me at one point, for which I am indebted.

Aiga Bautre, Ludmila Khvan, Uliana Kozhevnikova, Evgeni Magid and Tanya Tsoy, Mihail Pivtoraiko, Konstantin Salomatin, Jiangxia Shi, Maria Stankevich, George Skoptsov and Irina Zabbarova, Smita

Yadav, Yuki Yamamoto, Wanwan Zhang, WRCT staff and listeners, Klaus Sutner, Jon Uddstrom, and the CMU Aikido Dojo made Pittsburgh the most livable city for me.

Committing to a PhD program requires a belief that anything is possible, which I learned from my parents who have been supporting me in all my choices, all the way. My brother believed that anything is possible as long as he tries it first on me. To them, I miss you.

CONTENTS

1	INTRODUCTION	1
1.1	Designing a robot with ethnicity	2
1.1.1	The robot prototype hardware	2
1.1.2	Challenge 1. Identifying ethnic cues	4
1.1.3	Challenge 2. Believability of robot characters	8
1.2	Evaluating ethnic attribution and homophily	10
1.3	Terminology	10
1.4	Thesis contributions	12
1.5	Thesis overview	13
2	HOMOPHILY AND AGENTS WITH AN ETHNIC IDENTITY	15
2.1	Homophily in human-human and human-agent interaction	15
2.1.1	Gender	16
2.1.2	Personality	17
2.1.3	Ethnicity	17
2.2	Designing agents with ethnic identity	19
3	IDENTIFYING RICH POINTS	23
3.1	Related work	24
3.1.1	Nonverbal behaviors	24
3.1.2	Verbal behaviors	26
3.2	A Cross-Cultural Corpus of Receptionist Encounters	27
3.2.1	Data collection	27
3.2.2	Annotation scheme	30
3.2.3	Note on the analysis	31
3.2.4	Analysis: duration of interaction	31
3.2.5	Analysis: gaze	32
3.2.6	Smiles and nods	52
3.2.7	Lexical analysis	52
3.2.8	Conversation analysis	53
3.3	Summary	57
4	EVALUATING ETHNIC SALIENCE OF RICH POINTS	59
4.1	Crowdsourcing HRI	60
4.2	Evaluating verbal behaviors via text stimuli	61
4.2.1	Motivation	61
4.2.2	Procedure	61
4.2.3	Stimuli	62
4.2.4	Measures	65
4.2.5	Subjects	66
4.2.6	Results	67
4.2.7	Discussion	69
4.3	Joint evaluation of verbal and non-verbal behaviors	74
4.3.1	Selecting faces with high ethnic attribution	74

4.3.2	Scoring ethnic attribution of faces	77
4.3.3	Evaluating voices	83
4.3.4	Evaluating videos of verbal and non-verbal behaviors	86
4.4	Summary	93
5	ETHNIC BELIEVABILITY OF ROBOT CHARACTERS	95
5.1	Participants	95
5.2	Procedure	96
5.3	Stimuli	97
5.4	Measures	101
5.5	Results	102
5.5.1	Main effect of robot behaviors on ethnic attribution. Hypotheses 1AmE and 1Ar	103
5.5.2	Homophily effects. Hypotheses 2A-G	104
5.6	Discussion	107
5.7	Conclusions	109
6	TECHNICAL IMPLEMENTATION	111
6.1	The software architecture	111
6.1.1	Sensor and user input processing modules	111
6.1.2	Information retrieval modules	112
6.1.3	Expression system	112
6.1.4	Facial expression generation	112
6.1.5	Voice	113
6.1.6	Neck	113
6.2	Interaction manager	113
6.2.1	Related work	114
6.2.2	Overview	115
6.2.3	Interfacing with interaction manager	116
6.2.4	Interaction manager state	117
6.2.5	Control flow	117
6.2.6	Applications	119
7	SUMMARY AND FUTURE WORK	121
7.1	Summary of the results	121
7.2	Reflection	122
7.3	Future studies	126
7.4	Implications for technology	127
7.5	Applications in society	127
7.5.1	Cultural competence	127
7.5.2	Facilitating positive intergroup contact	128

APPENDIX 129

A	CROSS-CULTURAL PERCEPTION OF PERSONALITY THROUGH LANGUAGE	131
A.1	Stimuli	131
A.2	Results: Model fitting	139

B	EVALUATING VIDEOS OF VERBAL AND NON-VERBAL BE-	
	HAVIORS	145
B.1	Stimuli	145
B.1.1	Questionnaires	145
B.1.2	Greeting	150
B.1.3	Direction giving: Gaze	152
B.1.4	Handling failure to provide an answer	154
B.1.5	Handling disagreement	155
B.1.6	Direction giving: Politeness	155
B.1.7	User validation questions	155
B.2	Results	156
C	EVALUATING ROBOT BEHAVIORS	165
C.1	Stimuli	165
C.1.1	Protocol	165
C.1.2	Questionnaires	167
C.1.3	Locating the destination on the map	171
C.2	Results	172
C.2.1	Attribution of the robot characters as Ar	172
C.2.2	Attribution of the robot characters as AmE	173
C.2.3	Animacy of robot characters	174
C.2.4	Anthropomorphism of robot characters	175
C.2.5	Likeability of robot characters	176
C.2.6	Intelligence of robot characters	177
C.2.7	Safety of robot characters	178
C.2.8	Safety of robot characters (calibrated)	179
C.2.9	Locating the destination on the map	180
C.2.10	Thanking	180
	BIBLIOGRAPHY	181

LIST OF FIGURES

Figure 1	Hala at work. 2
Figure 2	Duration of interactions with receptionists grouped by native language and gender. Interactions where receptionists were not able to give any directions are excluded. 32
Figure 3	Duration of a receptionist's gaze on the visitor versus the length of the interaction. Numbers correspond to the participants' ids. The line corresponds to the best linear fit, $R^2 = 0.8$, $F(1,44) = 200.2$, $p < 0.01$. 33
Figure 4	Gaze on visitor and pointing gaze. 34
Figure 5	Duration of a receptionist's gaze on the visitor plus on the desktop versus the length of the interaction. The line corresponds to the best linear fit, $R^2 = 0.92$, $F(1,44) = 535.9$, $p < 0.01$. 35
Figure 6	Duration of a receptionist other gaze (including gaze on the desktop) versus the length of the interaction. 37
Figure 7	Gaussian kernel density estimate of female receptionist's continuous other gaze durations. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically. 38
Figure 8	Duration of a receptionist's gaze on visitor versus the length of the interaction. Interactions where receptionists were not able to give any directions are excluded. The data is grouped by the receptionist's native language and gender. 39
Figure 9	Duration of a receptionist's gaze on the visitor during the receptionist's speech. 40
Figure 10	Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English. 41
Figure 11	Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English. Female receptionists only. 41
Figure 12	Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English. Male receptionists only. 41

- Figure 13 Estimated distribution of durations of continuous gaze on visitor for female speakers of Arabic and American English. Only interactions where directions were given are shown. 42
- Figure 14 Duration of a receptionist gaze pointing versus the length of the interaction. As opposed to figure 4b, interactions where receptionists were not able to give any directions are excluded. Labels correspond to the full interaction code: the receptionist's id, the visitor's id and the id of the encounter for a given pairing of receptionist and visitor. 43
- Figure 15 Duration of a receptionist's gaze pointing versus the length of the interaction. Interactions where receptionists were not able to give any directions are excluded. The data is grouped by native language and gender. 44
- Figure 16 Distribution of durations of continuous pointing gaze for speakers of Arabic and American English. 45
- Figure 17 Distribution of durations of continuous pointing gaze for speakers of Arabic and American English. Female receptionists only. 45
- Figure 18 Distribution of durations of continuous pointing gaze for speakers of Arabic and American English. Male receptionists only. 45
- Figure 19 Gaussian kernel density estimates of female receptionist's durations of gaps in gaze on user and durations of continuous pointing gaze. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically. The dashed lines indicate the durations corresponding to the modes of the density estimates. 47
- Figure 20 Gaussian kernel density estimates of female receptionist's continuous pointing gaze durations with male and female visitors. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically. 48

Figure 21	Interaction between the visitor v1 and the receptionist r4. Wide vertical stripes represent intervals of speech. Narrow vertical stripes represent (from left to right): intervals of visitor's and receptionist's gaze towards the direction pointed by the receptionist, and visitor's and receptionist's gaze towards each other. Color coding of these modalities is specified by the icons in the upper part of the plots. 50	
Figure 22	Interaction between the visitor v11 and the receptionist r12. The visitor's eye gaze for this particular dialogue is partially inferred from his head gaze. Wide vertical stripes represent intervals of speech. Narrow vertical stripes represent (from left to right): intervals of visitor's and receptionist's gaze towards the direction pointed by the receptionist, and visitor's and receptionist's gaze towards each other. Color coding of these modalities is specified by the icons in the upper part of the plots. 51	
Figure 23	Receptionist r15 before and at the peak of lip stretch (AU20). Used with permission. 55	
Figure 24	An example web page with the naturalness item 65	
Figure 25	Selected results of verbal behavior evaluation 70	
Figure 26	Selected results of verbal behavior evaluation (continued) 71	
Figure 27	Selected results of evaluating the naturalness of verbal behaviors 72	
Figure 28	Human-like faces. 76	
Figure 29	Distributions of votes on ethnic attribution of the 18 human-like faces. The horizontal axis labels denote skin tone, eye, and hair color. 78	
Figure 30	The stimulus for evaluation of the ethnic attribution of a face. 79	
Figure 31	Human-like Face1 and Face2 were selected based on the large number of votes they collected as someone resembling native speakers of Arabic and American English, respectively. Face3 and Face4 are the robotic faces that elicit low ethnic attribution. 80	
Figure 32	Ethnic attributions of the four faces. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**). 81	

Figure 33	Ethnic attributions of the six Acapela voices. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**). 84
Figure 34	Experimental environment. 96
Figure 35	Score means on attribution of the robot characters as a native speaker of Arabic. Brackets correspond to 95% confidence intervals. This plot is for visualization only, as direct pairwise comparison would not account for subject effects. 104
Figure 36	Score means on attribution of the robot characters as a native speaker of American English. Brackets correspond to 95% confidence intervals. These plots are for visualization only, as direct pairwise comparison would not account for subject effects. 105
Figure 37	Conventional method for selecting behaviors of ethnic characters. 123
Figure 38	Proposed methodology for selecting behaviors of ethnic characters. 124
Figure 39	An example of a web page for one item of the experiment 138
Figure 40	A single stimulus web page presents the description of the situation, the video of the stimulus, items of the Godspeed questionnaire with their Arabic translations and the ethnic attribution items. Continued in figure 41. 146
Figure 41	Continued from figure 40. A single stimulus web page presents the description of the situation, the video of the stimulus, items of the Godspeed questionnaire with their Arabic translations and the ethnic attribution items. 147
Figure 42	A validation question for AmE participants. 148
Figure 43	A validation question for Ar participants. It asks "What did the robot say?" and provides a number of choices, only one of which is correct. 149
Figure 44	Task 1 interaction protocol that was given to participants. 165
Figure 45	Task 2 interaction protocol that was given to participants. 165
Figure 46	Task 3 interaction protocol that was given to participants. 166
Figure 47	The demographic questionnaire. 167

Figure 48	The baseline emotional state questionnaire. 168
Figure 49	Page 1 of the questionnaire: items of the God-speed questionnaire with their Arabic translations. 169
Figure 50	Page 2 of the questionnaire: ethnic attribution items with their Arabic translations, and the written part of map task. 170
Figure 51	The participants were given this map and asked to draw directions and mark the destination office. The destination labels (A for Adams, B for Brown, C for Coopers, and D for Douglas) were not shown to the participants. The location of the robot and the door to the lab where the experiment was conducted were included for reference. 171

LIST OF TABLES

Table 1	Distribution of participants between Doha and Pittsburgh experiment sites. The numbers of different participants in the receptionist role, whose interactions were annotated, are shown in parenthesis. 28
Table 2	Annotation scheme. 30
Table 3	Study participants by country 68
Table 4	The manipulated verbal and nonverbal behaviors. 98
Table 5	A fragment of codes for action units relevant for the expressions used in this study (based on FACS, Ekman and Friesen [1978]). 120
Table 6	Stimuli for the control condition 132
Table 7	Stimuli for the verbosity section 133
Table 8	Stimuli for hedging section of the experiment 134
Table 9	Stimuli for lexical and syntactic alignment section of the experiment 135
Table 10	Stimuli for formality section of the experiment 136
Table 11	Validation dialogue for AmE 137
Table 12	Validation dialogue for Ar 137
Table 13	Significant effects when varying verbosity 141
Table 14	Significant effects when varying hedging 142
Table 15	Significant effects when varying alignment 143
Table 16	Significant effects when varying formality 144

Table 17	Perception of greetings. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero. 157	
Table 18	Significant terms (up to two-way interactions) for scores of greeting stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1, neutral facial expression, and "Hi" are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero. 158	
Table 19	Perception of pointing gaze during direction giving. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero. 159	
Table 20	Significant terms (up to two-way interactions) for scores of direction-giving stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and gaze 1 are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero. 160	
Table 21	Perception of failures to provide an answer. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero. 161	
Table 22	Significant terms (up to two-way interactions) for scores of failure stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and explanation are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero. 161	
Table 23	Perception of disagreements. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero. 162	

Table 24	Significant terms (up to two-way interactions) for scores of disagreement stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and explicit disagreement are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero. 162
Table 25	Perception of politeness markers. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero. 163
Table 26	Significant terms (up to two-way interactions) for scores of politeness stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and direct style are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero. 163
Table 27	Significant associations with the robot characters' attribution as Ar by all participants. Interactions are denoted by colons. 172
Table 28	Significant associations with the robot characters' attribution as Ar by AmE participants. Interactions are denoted by colons. 172
Table 29	Significant associations with the robot characters' attribution as Ar by Ar participants. Interactions are denoted by colons. 172
Table 30	Significant associations with the robot characters' attribution as AmE by all participants. Interactions are denoted by colons. 173
Table 31	Significant associations with the robot characters' attribution as AmE by AmE participants. Interactions are denoted by colons. 173
Table 32	Significant associations with the robot characters' attribution as AmE by Ar participants. 173
Table 33	Significant associations with the animacy scores of robot characters by all participants. 174
Table 34	Significant associations with the animacy scores of robot characters by AmE participants. Interactions are denoted by colons. 174
Table 35	Significant associations with the animacy scores of robot characters by Ar participants. 174

Table 36	Significant associations with the anthropomorphism scores of robot characters by all participants. 175
Table 37	Significant associations with the anthropomorphism scores of robot characters by AmE participants. Interactions are denoted by colons. 175
Table 38	Significant associations with the anthropomorphism scores of robot characters by Ar participants. Interactions are denoted by asterisks. 175
Table 39	Significant associations with the likeability scores of robot characters by all participants. Interactions are denoted by colons. 176
Table 40	Significant associations with the likeability scores of robot characters by AmE participants. Interactions are denoted by colons. 176
Table 41	Significant associations with the likeability scores of robot characters by Ar participants. 176
Table 42	Significant associations with the intelligence scores of robot characters by all participants. 177
Table 43	Significant associations with the intelligence scores of robot characters by AmE participants. Interactions are denoted by colons. 177
Table 44	Significant associations with the intelligence scores of robot characters by Ar participants. 177
Table 45	Significant associations with the safety scores of robot characters by all participants. Interactions are denoted by colons. 178
Table 46	Significant associations with the safety scores of robot characters by AmE participants. Interactions are denoted by colons. 178
Table 47	Significant associations with the safety scores of robot characters by Ar participants. 178
Table 48	Significant associations with the calibrated safety scores of robot characters by all participants. Interactions are denoted by colons. 179
Table 49	Significant associations with the calibrated safety scores of robot characters by AmE participants. Interactions are denoted by colons. 179
Table 50	Significant associations with the calibrated safety scores of robot characters by Ar participants. 179
Table 51	Significant associations with thanking of robot characters by all participants. Interactions are denoted by colons. 180
Table 52	Significant associations with thanking of robot characters by AmE participants. Interactions are denoted by colons. 180

Table 53	Significant associations with thanking of robot characters by Ar participants. 180
----------	--

INTRODUCTION

More than machinery we need humanity.
— Charlie Chaplin, *The Great Dictator* (1940)

Humans tend to treat media, such as images, television sets, and computers, as social agents—they are polite, trusting and they bond with media in a similar way that they are polite, trusting and they bond with other humans. Reeves and Nass [1996] called this phenomenon the *Media Equation*, for the equality Media = Humans, since many of their studies show that substituting media for another human leads to similar results. Much evidence points that some of these mechanisms are subconscious. For example, study participants admit that being polite to a computer is not rational and they do not consider it as a living, sentient creature.

Capitalizing on the reliability of the Media Equation phenomenon, many contemporary media artefacts, such as computer interfaces and robots, are designed to elicit a desired social response. For example, an on-screen agent portraying a rude tutor may be more engaging to some students [Graesser et al., 2008]. A robot with its gaze and speech coupled to replicate a human speaker's behavior facilitates comprehension [Staudte and Crocker, 2009].

One desired social response is based on the pervasive phenomenon of similarity attraction, or *homophily* [Lazarsfeld and Merton, 1954]. Humans tend to relate with individuals who are similar along sociodemographic dimensions, such as gender, personality, or education. As the Media Equation predicts, homophily can be elicited between technology and its users. For example, Lee et al. [2000] showed that subjects conformed more to a computer voiced as the matching gender. Humans preferred the mechanical face that expressed the valence of extravertedness similar to their own [Park et al., 2012]. In the study by Nass et al. [2000], Korean students judged an embodied conversational agent (ECA, an on-screen agent that converses) with a Korean ethnic appearance as more trustworthy, attractive and competent, and also conformed more frequently to such agent's decisions as opposed to decisions made by a Caucasian agent.

While homophily along a variety of dimensions have been successfully elicited by both on-screen agents and robots, ethnic homophily



Figure 1: Hala at work.

has not been studied in human-robot interaction (HRI). This thesis addresses this void by developing a methodology for designing a robot with ethnic attribution and evaluating the homophily hypothesis.

1.1 DESIGNING A ROBOT WITH ETHNICITY

One of the reasons for the lack of studies on ethnic homophily in HRI is the difficulty of creating a robot with an ethnic identity. We contend that there are two main challenges. The first challenge, identifying ethnic cues, is a necessary step in the design of any ethnic agent, such as an ECA or robot. It requires bridging the gaps between the qualitative fields of cultural and linguistic anthropology, quantitative evaluation methods, and the technical implementation on a robot prototype. The second challenge, endowing a machine-like agent with ethnic attributes, is unique to physically embodied agents, or robots. It is closely related to the notion of believability. Below, we overview these challenges and our ways of tackling them, but first, we introduce the robot prototype.

1.1.1 *The robot prototype hardware*

We use the Hala robot receptionist prototype, introduced by [Simmons et al. \[2011\]](#), to generate videos of the behavior stimuli for an online study, as well as the actual robot prototype deployed in the main lab experiment described in this thesis. The robot consists of a human-like stationary torso with an LCD mounted on a pan-tilt unit PTU-46-17.5 manufactured by Directed Perception (now FLIR Motion Control

Systems). The LCD is used to render a 3D model of the robot's head, allowing for a relative ease in endowing the robot with both human-like and machine-like faces, as well as with faces of various ethnic appearances. The maximum torque limits on the PTU motors, however, restrict the maximum acceleration and, in the case of the tilt motor, also its range of the motion. This restriction affected the magnitude of the nod motion that we were able to implement on the physical head of the robot. To alleviate this problem, in addition to the expression of the physical nod, we also implemented an animated expression of an in-screen nod.

Hala is deployed around the clock as a robot receptionist (robot-receptionist) at the main lobby of the CMU Qatar campus. Figure 1 shows Hala interacting with a visitor, as the security staff and a human receptionist are at their duties nearby. Users interact with Hala, using a keyboard placed in front of her, in one of three available input modes: English, Arabic or transliterated Arabic (zarabi). There is an additional screen between the keyboard and the robot that displays the user's typed utterances and buttons that switch between the input methods. Depending on the input mode, Hala will respond in English or Arabic, by producing a synthesized voice reply as well as a text bubble that appears next to her face. Hala is designed to provide information about campus directions, weather, local events and answer queries regarding her personal life. Her backstory is that of a young unmarried Arab female robot. Her face, with its non-humanlike colors, is designed to minimize expression of ethnic cues via appearance. The robot uses a rule-based dialogue manager that consists of a knowledge base of canned utterances and rules that trigger responses in English and Arabic.

We used the Hala hardware to generate stimuli for online studies, as well as for a study with the robot colocated with the participants, described in this thesis. For ease of implementation of the stimuli behaviors, however, we used the next version of Hala's software architecture (Hala 2), with the technical contributions described in Chapter 6.

1.1.2 Challenge 1. Identifying ethnic cues

— Tashi . . . Your island culture is amazing!
 Even these juicy tropical candies!
 — Heather, I have a confession. These are Tropical Starburst.
 You can buy them anywhere.
 — Tropical Starburst, TV Ad by TBWA/Chiat/Day

1.1.2.1 Defining ethnic boundaries

Ethnicity

The factors that play a role in constructing a person's *ethnic identity* include language, religion, culture, appearance, ancestry, and regionality [Nagel, 1994]. However, as Waters [1990] notes, “ethnic identification is, in fact, a dynamic and complex social phenomenon.” In particular, individuals tend to invoke one ethnic identity or another depending on social context. For example, “someone whose mother is half Greek and half Polish and whose father is Welsh may self-identify as Greek to close friends and family and as Polish at work, or as Welsh on census documents” [Waters, 1990].

In addition to self-ascription, Barth [1969] distinguishes ascription of others as another critical feature of ethnicity. Charles E. Johnson [1974], for example, reports the results of the Current Population Survey by the US Bureau of Census, where interviewers were required to determine the race of those they were interviewing “by observation.” Out of total 22.9 millions of interviewees who were classified as “Negro” by the interviewer, 1.9 million answered something else for themselves.

Fluidity with which individuals ascribe ethnic identity to themselves and to others implies that any grouping based on static boundaries void of context is bound to have inaccuracies. In our task of identifying behaviors specific to a group, these inaccuracies may result in missing behaviors due to high intragroup diversity, or in finding behaviors that are salient for a superset of this group. We address this difficulty by validating the perception of behavior candidates as cues of ethnicity.

Specifically, we define two ethnic groups: *AmE*, native speakers of American English who are currently residing in the US, and *Ar*, native speakers of Arabic who are fluent in English and residing in one of the countries of Africa and Middle East that has Arabic as a main language. We relax the residency requirements to allow past residences for studies with physically present subjects (see the Section 1.3 for details). Each of these groups includes a number of distinct dialects and appearances. We address the intragroup variability by controlling for such sociodemographic features as gender, race, age, and countries of longest periods of residency.

Note that our definition of ethnic groups AmE and Ar does not involve race. In fact, we focus on behavioral rather than appearance cues of ethnicity, as we explain in Section 1.1.3.

1.1.2.2 Culture

— *You told me to take care of him.*
 — *Ah, shit! I meant to take care of him, not fuckin' take care of him!*
 — Frederick and Felix, *Formula 51* (2001)

Throughout this thesis we will prefer the term ethnicity to the term *culture* as less ambiguous. However, our definition of groups AmE and Ar in terms of native language rather than race, makes them closely related to the common use of the concept of culture. Frequently used in related work, the term *culture* is “one of the most widely (mis)used and contentious concepts in the contemporary vocabulary” [Agar, 2006]. Agar views culture in terms of the differences between a source language with its context (including the situation and the interlocutor’s background) and a target language with its context. The combination of language and its sociocultural context is referred to as *languaculture* by Agar [1994]. As an example, Agar describes a junkie languaculture, *junky* being the term drug addicts use to refer to themselves. Their vocabulary (e. g. “get off”), which is distinct from standard English, is tied to their situations (e. g. shooting heroin), that are distinct from the situations of non-addicts.

Agar refers to such differences, that are in the center of the definition of culture, as *rich points*. In ethnographic work, rich points are discovered when the ethnographer’s expectations differ from what he observes. For example, a university faculty member in the US may find it unusual the first time a foreign student addresses her as “professor.” The term of address would be a rich point between the professor’s and the student’s ways of using the language in the context. Note that this rich point can be a cue to the professor that the student is a foreigner, but may not be sufficient to further specify the student’s ethnic identity.

This thesis addresses this apparent difference between the rich points and ethnic cues by evaluating salience of the rich points as ethnic cues as perceived by the target community.

Culture

Languaculture

Rich point

Rich point vs. ethnic
cue

1.1.2.3 *Selecting ethnic cues*

Lord Helmet: Found anything yet?
White troopers: Nothing yet, sir!
Lord Helmet: How about you?
Other white troopers: Not a thing, sir!
Lord Helmet: What about you guys?
Black troopers: We ain't found shit!
 — *Spaceballs* (1987)

Quantitative analogs of rich points, referred to as *maximally distinctive behaviors* by [Iacobelli and Cassell \[2007\]](#), are commonly used as a proxy for ethnic cues in the design of on-screen characters with an ethnic identity. For example, [Iacobelli and Cassell \[2007\]](#) observed that African American children who were monolingual speakers of African American Vernacular English gazed more at the toys than one another, as opposed to Caucasian children who spoke Standard American English. Such behaviors are discovered through analysis of human interaction corpora, making them potentially a more objective description of differences between ethnic groups as compared with rich points, that are dependent on the identity of the ethnographer. However, just like rich points discovered through ethnographies, maximally distinctive behaviors are not necessarily salient ethnic cues (see, for example, [\[Iacobelli and Cassell, 2007\]](#), for the discussion on the difficulty of evoking an African American attribution through behaviors of an on-screen character).

Nevertheless, we consider both rich points discovered through ethnographies, and maximally distinctive behaviors found by corpora analyses, as viable candidates for ethnic cues. We, however, augment the traditional methodology of building an on-screen agent with an ethnic identity by performing an additional evaluation of ethnic salience of the candidate behaviors. For brevity, unless there is a need to highlight the difference, we will use the term rich points to refer to both the results of qualitative ethnographies and maximally distinctive behaviors found quantitatively.

As we explain in Section 1.1.3, we start with rich points between ethnic groups AmE and Ar that are expressed as they converse in English. Rich points can make their way into all linguistic levels of a second language. For example, they can be a cause of *pragmatic failure*, a topic that has received a qualitative account in the literature (see, for example, [\[Aguilar, 1998\]](#) for an overview). The few known quantitative accounts of rich points are limited in the scope of behaviors and contexts across which behaviors are compared (see, for example, [\[Schneider, 2007\]](#) for a comparison of English thanking sequences across regions and speech genres of shopping encounters and radio

interviews). We review the efforts on identifying maximally distinctive behaviors, as a step in the process of creating on-screen agents with ethnic identity, in Section 2.2.

In this work we propose an extra step of evaluating the salience of the candidate behaviors as ethnic cues via crowdsourcing. In addition to the argument that differentiates rich points from ethnic cues that we gave earlier, such an evaluation allows one to pretest a wider range of behaviors via an inexpensive, online study, before committing to a final selection of ethnic cues for a more costly experiment with a physical robot colocated with the participants. The augmented methodology of creating an agent (a robot or an on-screen character) with an ethnic identity is as follows:

- Initial candidates are selected from ethnographies (for example, from known rich points), quantitative human-human corpora analyses (as maximally distinctive behaviors), research literature on second language acquisition, ECAs, and other areas;
- Candidate behaviors are rendered on-screen and evaluated quantitatively online via crowdsourcing (e.g. Amazon’s Mechanical Turk);
- Finally, the most salient ethnic behaviors are implemented in a robot prototype and user studies are conducted.

For example, the following verbal and non-verbal behaviors are potentially rich points between native speakers of American English and Arabic.

- Feghali [1997] summarizes the following verbal behaviors shared by native speakers of Arabic: repetition, indirectness, elaborateness, and affectiveness. The two questions relevant for our study are: (1) to what extent are these features transferred to native speakers of Arabic speaking English, and (2) what are the linguistic devices that realize these features.
- Non-verbal behaviors that are candidate rich points include eye gaze patterns, such as gaze and glance durations. For example, Leathers [1997] describes differences between direct eye gaze behavior among Arabs and Japanese. In general, however, there is little data about differences between non-verbal behaviors of native speakers of Arabic and American English.

We evaluate these and other candidate rich points by first confirming their occurrences in the corpus of human-human interactions (Chapter 3), and then measuring their perceived qualities, including ethnic attribution, by members of AmE and Ar via crowdsourcing (Chapter 4). Finally, we implement the behaviors on a robot prototype and evaluate ethnic attribution and homophily via a controlled

study with the participants colocated with the robot (Chapter 5). We found some evidence of ethnic attribution, but no strong evidence of homophily. There is, however, one more challenge that a designer of an ethnically salient robot character must face. We describe it in the following section.

1.1.3 Challenge 2. *Believability of robot characters*

Andy Serkis and the other actors playing apes could be put on regular sets but they did have to wear gray pajamas with wires all over them and you know these little arms that came out from their faces with cameras pointed at their faces to capture their expressions. And you'd think that acting opposite someone in that kind of outfit would pull you out of the imaginary circumstances. How could you ever believe that that was an ape?! But their behavior was so believable that almost immediately all of that stuff melted away and you could just interact with them as chimpanzees and really let the imagination take over, 'cause they were so good at it. That truly was an enjoyable experience.

— James Franco on *Rise of the Planet of the Apes* (2011)

In spite of the difficulties of identifying ethnic cues that we outlined earlier in Section 1.1.2, there has been considerable work designing ethnic on-screen agents (see Section 2.2 for an overview). Why is designing an ethnic robot a harder problem than designing an ethnic on-screen agent? We think that the answer is in the notion of believability.

Believability

For the Media Equation to work, humans need a degree of a suspension of disbelief that can create an illusion of life, referred to in the arts as *believability* [Thomas and Johnston, 1981; Bates, 1994]. Believability of a character is an objective in literature, theater, animation, film and other media. It is associated with engagement and emotional empathy with the character's emotions and predicaments as experienced by the members of the audience.

While ECAs may be portraying human characters much like it would be done in an animated film or a puppet theater, we contend that with few exceptions contemporary robots retain their machine-like qualities and their embeddedness in the real world (or, in theatrical terms, “broken fourth wall” [Bell, 2008]). The first of these factors is due to mechanical limitations of robots that do not apply to ECAs. While there is some work on translating principles of animation to robot movement (for example, the recent work by Ribeiro and Paiva [2012]) as well as work on human-like robots with extremely realistic appearance (such as Geminoids by Ishiguro [2005]), most of the

contemporary robots easily give away their mechanical nature either by their appearance, or, more certainly, by their movement. The second factor, “broken fourth wall,” is a consequence of the fact that with a few exceptions (for example when robots perform on stage, c.f. Hayashi et al. [2005]; Lin et al. [2009]; Ogawa et al. [2012]), robots interact with people in a shared physical space, making it harder to suspend disbelief.

Expressing ethnic cues with on-screen agents is too not completely problem-free. For example, behaviors that cater to stereotypes of a certain community can be found offensive to the members of the community the agent is trying to depict [de Rosis et al., 2004]. We speculate that endowing low-believability robots with strong ethnic cues, such as appearance, may lead to similar undesirable effects. For example, from the conversations the author conducted on this issue in Qatar, people are split on their opinions about a possibility of Hala wearing traditional items of clothing, such as *abaya*. One reason is that traditional clothing is associated not just with ethnicity, but with a multitude of other dimensions of sociodemographic identity, such as religious attitudes.

In this work, we address this challenge by focusing on subtle behavioral cues, beyond appearance and the choice of speaking Arabic or English. These cues include verbal behaviors, such as politeness strategies, and non-verbal behaviors, such as gaze patterns. Although such behaviors may have smaller effect on ethnic attribution, there are several arguments in support of using subtle cues of ethnicity:

- In multi-party interactions in a multi-cultural environment, having a robot switch to the native language of a particular user may lead to undesirable effects of exclusion of other participants.
- For any conversational agent, the choice has to be made on whether a particular verbal and non-verbal behavior should be expressed. Knowing how much subtle behaviors contribute to ethnic attribution allows for an informed choice that optimizes expression of a particular ethnicity.
- As argued by Iacobelli and Cassell [2007], “physical appearance is not the most reliable index of ethnicity.” This is especially true for our particular definition of ethnic groups AmE and Ar. In this thesis, we validate this statement by evaluating ethnic attribution of images of human-like and robotic faces that we then use in interaction studies to control for physical appearance.
- We contend that subtle behaviors will not trigger the potential offensiveness associated with a robot expressing ethnic cues.

For these reasons, we focus on verbal and non-verbal behaviors that are expressed by members of AmE and Ar when speaking English.

Why focus on subtle behavioral cues of ethnicity?

1.2 EVALUATING ETHNIC ATTRIBUTION AND HOMOPHILY

Once we identify behavioral cues of ethnicity via corpus analysis and crowdsourcing experiments, we incorporate them into the behaviors of a robot prototype and conduct the laboratory experiment with physically present subjects. Our goal is to evaluate two families of hypotheses. The first family is concerned with our ability to elicit ethnic attribution by varying the behaviors of the robot characters. The second family is concerned with the effect of ethnic congruence between the robot character and the participants on perceptual and task performance measures.

The subjects engage in a realistic direction-seeking dialogues with several robot receptionist characters that vary with respect to their behavioral and, for control, appearance cues of ethnicity (faces). After several interactions with a robot character, the participants evaluate the character's ethnicity and other perceptual measures via questionnaires. They are also asked to locate the destination of the provided directions on a map. The performance on this task, intended to measure comprehension and recall, and the number of times the participants thanked the robot, serve as objective measures of homophily. The procedure repeats until the participant evaluates four different robot characters.

A peek into results.

We analyze the data by performing hypothesis testing on fitted linear models. We find partial support for the attribution family of hypotheses. In particular, the robot's behaviors affect perception of the robot characters as native speakers of American English by female participants. However, the robot's behaviors affected the perception of the robot character as a native speaker of Arabic only for two of the four faces.

Our analysis does not find support for any of the homophily hypotheses. For discussion of possible reasons and implications we refer the reader to Section 5.6.

1.3 TERMINOLOGY

In the section, we define the main terms that we will use throughout this document. Many of these definitions use other terms defined in this section. Nevertheless, there are no directed cycles.

AmE. Native speakers of American English (*L1* American English speakers). For online studies, we further restrict this group to those currently residing in the US. For studies with physically present subjects, we require either at least one year of prior residency in the US, or prior attendance at an American international school. AmE defines an ethnic group and a corresponding languaculture (see below).

Ar. Native speakers of Arabic (*L1* Arabic speakers). For online studies, we further restrict this group to those who currently reside in one

of the countries of Africa or the Middle East that has Arabic as an official language. These countries are: Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestinian Territories, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen. We exclude countries where Arabic is an official language but is spoken by a minority of the population, such as Chad, Comoros, Djibouti, Israel, and Somalia. For studies with physically present subjects, we require at least one year of prior residency in one of those countries. Ar defines an ethnic group and a corresponding languaculture. Note, that we will study the behaviors that are expressed by Ar when speaking English as a second language (*L2*).

Culture. Following Agar [2006] we view culture in terms of the differences between a source languaculture and a target languaculture. In relation to our experiments, we will sometimes refer to ethnically salient behaviors as belonging to a certain languaculture. Such behaviors are often referred to as culture in the works of other authors. Sometimes we will stick to our terminology even when describing their work, and will attempt to clarify the difference when there is a chance of confusion.

Ethnic congruence. Similarity of ethnicities, ethnic attributions, ethnically salient behaviors, or ethnically prevalent behaviors between two interlocutors.

Ethnicity. Ethnicity, or *ethnic identity* of a person is constructed by the person or by others based on factors that include language, religion, culture, appearance, ancestry, and regionality [Nagel, 1994]. Ethnic identity is a fluid concept that may change with time and context. In our work, we attempt to overcome this fluidity by defining ethnic boundaries in terms of native language and countries of current or prior residency (refer to the definitions of AmE and Ar in this section). Behaviors that are prevalent within ethnic groups and different across them are referred to as *rich points* (see below). Behaviors that evoke an attribution of ethnicity are *ethnic cues* (see below).

Ethnically distinctive behaviors. Behaviors that differ between two ethnic communities. We will overload the notion of *rich points* to include such distinctive behaviors. Not that not all ethnically distinctive behaviors are ethnic cues.

Ethnic salience (ethnic cue). We refer to a behavior as ethnically salient (or as an ethnic cue) if it can evoke an ethnic attribution by a member of a target community. In other words, in this work, we say that a behavior is salient of ethnicity A to members of ethnicity B; or that a behavior is an cue of ethnicity A to members of ethnicity B.

Homophily. A tendency of individuals to associate and bond with similar others.

Interlocutor. A participant in a conversation.

Languaculture. Languaculture is language with its context, including the situation and the interlocutor's background. In relation to our

experiments, we will sometimes refer to behaviors as being specific to a certain languaculture (AmE or Ar). Since we follow the view of culture in terms of the differences of languacultures, namely in terms of rich points (defined below), we will use the term languaculture where some other authors would use the term culture. Note, that in case of the languaculture Ar, we will study the behaviors that are expressed when speaking English as a second language.

Rich point. Rich points are behaviors that signal the differences between languacultures [Agar, 1994]. We will also use this term in a broader sense, to refer to distinctive behaviors between two ethnic groups. We distinguish rich points from ethnically salient behaviors, as not every difference between ethnic groups is a cue of ethnicity. For example, a Cantonese sentence ending particle “lah,” often transferred into L2 English, is a rich point between American English and Cantonese languacultures. However, unless the listener is familiar with this fact, hearing “lah” may not necessarily evoke attribution of the speaker as a native speaker of Cantonese.

Robot character. In earlier work, we developed the notion of a *believable robot character* as a performer with a cohesive backstory, behaviors, and appearance [Simmons et al., 2011]. In this thesis, we are concerned with robots that use appearance and behaviors to express their sociodemographic identity, in particular, ethnicity. Throughout the thesis, we will be referring to the combination of a robot’s appearance and behaviors that aims to express a sociodemographic identity as a *robot character*.

1.4 THESIS CONTRIBUTIONS

In this thesis, we advocate for an explicit treatment of ethnic identity in the design of believable robot characters. The main contributions of this work are as follows.

- Methodology for identifying salient ethnic behaviors via on-screen simulations and crowdsourcing. We combine these behaviors to create robots that evoke ethnic attribution. While this thesis focuses on robot characters, this methodology can be potentially applied to the design of behaviors of on-screen agents.
- Evaluation of effects of human-robot ethnic congruence on interactions.

The main hypotheses of this thesis are as follows:

Hypothesis I (Ethnic attribution) Believable ethnic identity of robot characters can be created using behaviors selected via lower fidelity on-screen simulations and crowdsourcing.

Hypothesis II (Homophily) Human-robot ethnic congruence improves subjective and objective measures of interactions.

1.5 THESIS OVERVIEW

The remainder of the document is as follows. Background and related work is reviewed in Chapter 2. Chapter 3 describes our work identifying culturally salient behavior candidates from existing studies and our own corpus of service encounters. Chapter 4 presents our results in evaluating ethnic salience of verbal and non-verbal behavior candidates. In Chapter 5, we use selected behaviors to create ethnically salient robot characters and evaluate their ethnic attribution and homophily in a controlled lab experiment. The technical implementation of this work is described in Chapter 6. Finally, Chapter 7 summarizes the results, and outlines directions of future work.

HOMOPHILY AND AGENTS WITH AN ETHNIC IDENTITY

We start the chapter with a review of work on homophily in human-human and human-agent interaction through similarity via the dimensions of gender, personality, and ethnicity. We then describe examples of agents (ECA and robots) with ethnic identities and position our own work in this space.

2.1 HOMOPHILY IN HUMAN-HUMAN AND HUMAN-AGENT INTERACTION

*The insiders in a we-group are in a relation of
peace, order, law, government, and industry, to each other.
Their relation to all outsiders, or others-groups, is one of war and plunder....
Sentiments are produced to correspond.
Loyalty to the group, sacrifice for it, hatred and contempt for outsiders,
brotherhood within, warlikeness without—
all grow together, common products of the same situation.*

— William Sumner, *Folkways* (1906)

Human social networks display homophily effects along a diverse set of dimensions, from the easily accessible individual characteristics, such as demographic factors of age, gender, ethnicity, religion, education and occupation (see [McPherson et al. \[2001\]](#) for an overview), to deeper individual characteristics, such as personality (e.g. [Selfhout et al. \[2010\]](#)). The effect of homophily is also observed in arbitrarily defined dimensions of similarity. Namely, randomly categorizing people into groups can lead members to perceive outgroup members as less trustworthy, honest, and cooperative than members of their own group, even without any interaction between or within the groups (see [Brewer \[1979\]](#) for an overview of research on factors affecting ingroup bias).

The study of homophily between humans and agents has a shorter history, with fewer dimensions of similarity explored and less agreement about the effect of similarity on performance. In this section, we review work on human-human and human-agent homophily induced via similarity of gender, personality or ethnicity.

2.1.1 Gender

Sex and gender homophily begins from an early age, with school children being more likely to resolve intransitivity of friendship relations (namely, when A likes B, and B likes C, but A does not like C [Hallinan and Kubitschek, 1990]) by deleting a cross-sex friendship than by adding another cross-sex friendship [Tuma and Hallinan, 1979]. In fact, Tuma and Hallinan found that most children are more likely to delete a same-sex friendship than to resolve the intransitivity by adding a cross-sex one. These early inbreeding processes contribute to development of adult gender homogeneity of both work establishments and voluntary associations (see McPherson et al. [2001] for an overview). This gender homogeneity of real world work groups sharply contrasts with studies that show that gender-heterogeneous teams outperform homogeneous ones in many business scenarios (see, for example, Hoogendorn et al. [2011]; Apesteguia et al. [2012]). Gender preferences in service encounters display sensitivity to interaction scenarios. For example, patients generally prefer same-sex physicians (e.g. Levinson et al. [1984]). On the other hand, men buying a massage service for no specific health reason tend to prefer a masseuse, while the women's choice is more mixed and is conditional on a number of factors [Poria, 2008].

One of the early experiments with gendered computer-generated speech showed that subjects liked and conformed more to a male-voiced computer but, controlling for this main effect, subjects conformed more to the computer voiced with a matching gender [Lee et al., 2000]. In a followup study that used a cartoon character and controlled for the gender stereotype of the task (sports versus fashion), Lee [2003] showed, however, that subjects' conformity to the computer agent is associated not with the interaction between the genders but rather with the pairwise interactions between the subject gender and the task and between the agent gender and the task.

Siegel et al. [2009] studied the persuasiveness of a museum robot that varied its voice between pre-recorded human female and male voices. They found cross-gender preference on measures of credibility, while men rated the robot with the female voice as more trustworthy and engaging. Men tended to donate more to the female-voiced robot, while women donated more to female-voiced robot when they were accompanied by another person.

2.1.2 *Personality*

— *Why don't you give me a little bit on personality traits that you are looking for.*

— *Yes, I want her to be blond.*

— Borat, *Da Ali G Show* (2003)

A study of friendship networks among just-acquainted late adolescents by Selfhout et al. [2010] showed that individuals tend to select friends with similar levels of agreeableness, extraversion and openness. A review of studies on personality composition of workgroups, conducted by Halfhill et al. [2005], highlighted associations between the group's homogeneity with respect to personality traits and the group's performance. Halfhill et al. blame *personality clash* [Vaccaro, 1988] as one of the reasons of the performance decrease.

Studies of the relationship between personality dimensions and attraction of ECAs and robots have mixed results. Isbister and Nass [2000] studied a full-body ECA that expressed extravertedness dimension of personality with both verbal and non-verbal behaviors. They found that participants tended to prefer characters with personality that is complementary, rather than similar, to their own. On the other hand, using a mechanical face to express only non-verbal behaviors, Park et al. [2012] showed that participants preferred the robot with a similar valence of extravertedness. The participants did, however, perceive the robot with complementary personality as more socially present. Lee et al. [2006] reported complementary preference for the valence of extravertedness expressed via verbal and non-verbal behaviors of a Sony AIBO. Tapus et al. [2008] found that an assistive therapy robot that matched the valence of extravertedness of stroke patients via verbal and non-verbal behaviors was successful in increasing the durations of therapeutic interactions. However, robots are not always easily attributed personality. For example, Woods et al. [2005] report that personality attribution to a non-anthropomorphic robot is positively associated with technological experience and negatively associated with the participant's age.

2.1.3 *Ethnicity*

Individuals who share similar demographic characteristics are drawn to each another, as similarity provides a source of familiarity, predictability, comfort, and validation (see, for example, Williams and O'Reilly [1998] for a review).

Studies of human workgroups show mixed relationship between cultural diversity and workgroup outcomes (see, for example, Barinaga [2007] for an overview). For example, Watson et al. [1993] showed

that culturally homogeneous (white) student work teams had better interaction and performed better on tasks of case analysis than culturally heterogeneous (white, black, Hispanic, Asian and Middle-Eastern) work teams. However, the so-called paradox of heterogeneity presented by Blau [1977] suggests that national diversity can weaken barriers instead of heighten barriers to effective performance in groups. Thus, an analysis of performance of NHL teams in terms of winning percentage reveals a U-shaped relationship between team national heterogeneity and team performance [Phillips and Phillips]. In other words, teams that are low and high on national heterogeneity outperformed teams that are in-between.

Congruence

To explain these inconsistent findings, Ely and Thomas [2001] and Barinaga [2007] suggest that deeper factors may be at play, such as discourse about national identity. Behrend and Thompson [2011], for example, refer to similarity expressed via appearance as shallow-level similarity, as opposed to deep-level similarity achieved via matching behaviors and attitudes. In fact, expressing *congruence* (similarity) via ethnically specific non-verbal behaviors is shown to be a better predictor of attraction than ethnic congruence itself [Dew and Ward, 1993]. In particular, Dew and Ward asked Palagi (white New Zealand) women to interview one of two female confederates, who was either Palagi or Samoan. Each of the confederates could express Palagi and Samoan non-verbal styles, enabling 2×2 between-subject design: confederate ethnicity \times confederate non-verbal style. In the authors' words: "Findings revealed that nonverbal style, but not ethnicity, significantly affected person perception in intercultural encounters. More specifically, subjects preferred individuals who exhibited culturally congruent nonverbal behaviors." The fact that ethnic congruence may not be necessary for ethnic homophily suggests a possibility that a non-humanlike agent that is not capable of evoking an ethnic attribution may still elicit a positive response by using congruent ethnic behaviors.

Ethnic congruence in agents, namely the property when the ethnic identity expressed by an agent coincides with the ethnicity of a user, is known to improve interactions. For example, Baylor and Kim [2003] showed that African-American students may affiliate more strongly with an African-American agent, and perceive it as more engaging and more facilitating of learning. A later study by Baylor and Kim [2004], however, showed that both black and white students learned more with a black agent. Nass et al. [2000] compared a Korean and a Caucasian agent, where the difference was expressed via appearance only, and found that participants of the same ethnic group judged the congruent agent as more trustworthy, attractive and competent, and also conformed more frequently to the congruent agent's decisions. Iacobelli and Cassell [2007] studied story telling with Virtual Peers (see Section 2.2) that expressed their ethnicity via both verbal and

non-verbal behaviors. They found a trend where African American children would tell back longer stories to the agent with congruent ethnicity.

In the previous paragraph we outlined several studies that evaluate homophily between users and on-screen agents. To the best of the author's knowledge, ethnic homophily effects have been not tested with robot characters. In the following section, we focus on the aspects of design of an agent with ethnic identity.

2.2 DESIGNING AGENTS WITH ETHNIC IDENTITY

When in Rome, they do as he does.

— The Most Interesting Man in the World,
Dos Equis TV ad by Euro RSCG

The majority of studies that involve ECAs with ethnically-authentic traits model ethnicity through appearance of the agent and its language choice (for example, Spanish versus English). Similar bias towards appearance-based expression of ethnicity is observed in robot characters, such as Ibn Sina [Mavridis and Hanson, 2009] and androids [Ishiguro, 2005]. On the other hand, a complete character design approach advocated by Hayes-Roth et al. [2002] suggests that ethnic, as well as individual variability, should be expressed in all of “the ten key characteristic qualities that animate characters—distinguished from other synthetic characters by their lively autonomy and individual personas—should possess.” These ten qualities, according to Hayes-Roth et al. [2002], are identity, backstory, appearance, content of speech, manner of speaking, manner of gesturing, emotional dynamics, social interaction patterns (e.g. how does the character address and react to other interactors depending on their gender, age and status), role and role dynamics.

Recent years saw the development of ECAs that express their ethnic identity via some of these qualities, beside the appearance and the choice of language. Many of them rely on the notion of *national culture*, formalized by Hofstede [2001] via five dimensions of variability:

Hofstede's model

1. *Small versus large power distance*, or “the extent to which the less powerful members of organizations and institutions (like the family) accept and expect that power is distributed unequally.”
2. *Individualism versus collectivism*, or the degree to which individuals in the society are integrated into groups.
3. *Masculinity versus femininity*, or “how central a role traditionally male values like earnings, recognition, advancement and challenge play in a society.”

4. *Weak versus strong uncertainty avoidance*, i.e. “the extent to which uncertain or unknown situations are seen as a threat.”
5. *Long versus short term orientation*. Long term orientation is association with such values as thrift and perseverance. Short term orientation is associated with respect for tradition, fulfilling social obligations, and protecting one’s “face.”

In spite of the criticism, Hofstede’s model of national culture can provide some intuitions about possible rich points. In Section 3.1.1, for example, we will use the difference in power distance scores between a number of predominantly Arabic-speaking countries and the US to speculate on the possible attitudes towards unequal status encounters and, as a result, ethnic differences in gaze behaviors of a receptionist. For now, we continue with the review of several advanced ethnic ECAs.

Kyra (Kira, Kirita)

The Kyra (Kira, Kirita) character, is an adolescent girl from the US (Brazil, Venezuela) [Hayes-Roth et al., 2002], who both speaks in the appropriate dialects of American English (Brazilian Portuguese, and presumably, Venezuelan Spanish), and also chooses ethnically-specific conversation content (*shibboleths*) and displays localization-dependent degrees of shame when discussing the fact that she is an orphan. Some of the ethnic traits implemented by Hayes-Roth et al. [2002] appear to be derived from speculations on values along the Hofstede-like dimensions of national culture, such as individualism and power distance; others have no clear indication of their sources.

Virtual peers

Iacobelli and Cassell [2007] describe an African-American and Caucasian-American virtual peers that interact with children using elements of either African American Vernacular English (AAVE) or Standard American English (SAE), have different eye gaze patterns, but share the same appearance. African-American children who participated in the study showed a trend towards perceiving the agent’s ethnicity and mimicked the SAE virtual peer’s linguistic behaviors more frequently. Verbal and non-verbal behaviors in this study were identified via corpus analysis of interactions between dyads of Caucasian and African-American children. We deploy a similar approach by analyzing the videos of interactions between Ar and AmE in Section 3.2.

African American children were able to more readily recognize the ethnic attribution of the agent with Caucasian behaviors, rather than the ethnic attribution of the agent with African American behaviors. This result is somewhat similar to the result of our final experiment, where we were more successful in eliciting attribution of the robot as

a native speaker of American English, rather than a native speaker of Arabic (see Section 5.5.1).

Carmen

Yin et al. [2010] introduce Carmen, an ECA that attempts to promote healthy habits among older users by building long-term relationships. The adaptation of the agent from an Anglo-American to a Latino-American population involved (a) changes in language of power-distance, namely, appeals to authority “were reworded or removed to increase the counselor’s affinity,” [Yin et al., 2010] and (b) appeal to collective activities, to reflect the findings from Hofstede’s cultural dimensions research. The authors note the difficulties of targeting a particular culture, including the lack of access to the target user population and lack of expertise in the target culture by the development team. The first part of this thesis attempts to address these issues by using crowdsourcing (see Chapters 3 and 4).

CUBE-G

The German-Japanese project CUBE-G is one of the most extensive data-driven efforts of identifying and generating ethnically-specific behaviors. [Rehm et al., 2009b] describes how analysis of body movements of Japanese and German participants during interactions with actor confederates was used to create a model that relates parameters of non-verbal behaviors and the values of Hofstede’s cultural dimensions. Specifically, Rehm et al. [2009a] collected a cross-cultural multimodal corpus of dyadic interactions in three scenarios (first encounter, negotiation, and interaction with an individual of a higher status), that they annotated with respect to head, arm, and leg postures. They also coded gestures on the expressivity scale adopted from Pelachaud [2005]. Finally, they trained a Bayesian network that connects Hofstede’s cultural model parameters with the parameters of posture and expressivity scales.

To the best of the author’s knowledge, there has not been a robot character with a comparably explicit treatment of its ethnic identity. This thesis addresses this gap.

IDENTIFYING RICH POINTS

The primary goal of the study described in this chapter is to select viable candidate behavioral cues of native speakers of American English (AmE) and native speakers of Arabic, as they are speaking English as a foreign language (Ar). In particular, we are focusing on a specific type of service encounter: the interactions with a receptionist. Similarly to the common approach of designing agents with ethnicity, we start with identifying rich points and maximally distinctive behaviors from qualitative and quantitative studies. We focus on behaviors that are feasible in the Hala robot prototype, introduced in Section 1.1.1. For example, its inability to move its base and torso, or to perform translational motions with its head implies that we cannot explicitly control interpersonal distance. We also exclude possible prosodic differences for the sake of keeping the scope of the work within reach. Our primary focus among non-verbal behaviors is gaze, head gestures such as nods and head shakes, and facial expressions, such as smile. For the verbal behaviors, we are looking for differences in lexical choices, syntactic structure as well as pragmatic strategies (e.g. giving an excuse).

First, we review the existing work on rich points between AmE and Ar groups along these verbal and non-verbal dimensions. It is generally agreed that many of these behaviors are highly dependent on the context. Since there are no studies of receptionist encounters, we collected our own video corpus of such encounters, annotated and analyzed it. This section describes this data collection and analysis and ends with a discussion of the candidates we select for the evaluation of their ethnic salience by the studies described in Chapter 4.

3.1 RELATED WORK

3.1.1 *Nonverbal behaviors*3.1.1.1 *Gaze and status*

Не смею требовать любви.
 Быть может, за грехи мои,
 Мой ангел, я любви не стою!
 Но притворитесь! Этот взгляд
 Всё может выразить так чудно!
 Ах, обмануть меня не трудно!..
 Я сам обманываться рад!

*I do not dare to plead for love;
 Love, for the sins I have committed,
 I am perhaps unworthy of.
 But make believe! Your gaze, dear elf,
 Is fit to conjure with, believe me!
 Ah, it is easy to deceive me...
 I long to be deceived myself!*

— Alexandr S. Pushkin, *Confession* (1826)
 Translated by Katharena Eiermann

Amount of gaze on the interlocutor is associated with *status* (see [Ridgeway et al., 1985] for an overview), where status can be defined as “having control or influence over another or possessing privileged access to restricted resources” [Mast and Hall, 2004]. Several studies suggest that Arabic-speaking societies may have a different attitude towards status than, for example, the United States. Nydell [1987], for example, emphasizes the importance of family background and social class in determining personal status in Arab societies (as opposed to individual character or achievement). Hofstede’s model of national culture, introduced in Section 2.2, places “Arab world” (the term he uses for Egypt, Iraq, Kuwait, Lebanon, Libya, and Saudi Arabia) at the high end of power distance spectrum, while the US is at its lower end [Hofstede, 2001]. Since power distance is defined as “the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally,” it is likely to be related with attitudes towards status. In the context of our study, we can speculate a more significant status difference between the receptionist and the visitor in an Arabic-speaking society, as opposed to the US.

How does the status differences affect gaze? Exline et al. [1975] note that peers look more at their partners while listening than while speaking, while high-status male ROTC officers looked at cadets about

the same amount of time while speaking as while listening. Low-status cadets looked substantially more while listening than while speaking. Ellyson et al. [1980] confirms this pattern among female dyads. This pattern, however, may not be universal, as Johnson [1976] reports that African Americans tend to look at their conversation partner more while speaking than while listening. It has also been suggested by Argyle [1967]; Strongman and Champness [1968] that breaking first the initial eye contact is a nonverbal sign of deference or submission.

If our hypothesis that status difference would be more apparent in the behaviors of a receptionist and a visitor in an Arabic-speaking society is true, and their behaviors in the unequal status encounters are similar to those described in the cited studies, we could expect Arabic receptionists to exhibit more gaze on visitor during their own speech as compared with American receptionists.

3.1.1.2 *Head movement*

Head movement has been studied in both the context of conversation (e.g. by McClave et al. [2007]) and as emblematic gestures that can be understood without speech (e.g. by Safadi and Valentine [1990]). Safadi and Valentine note that side-to-side shake signaling negation in American culture is now also used to some extent by Arabs, especially in Saudi Arabia. However, they claim that upward toss of the head is still a common emblem for “no” in Arab cultures. As for the head movements that co-occur with speech, McClave et al. [2007] shows that native speakers of Arabic, Bulgarian, Korean and African American English use the same patterns of head movements to express inclusivity, lists of alternatives and deixis (i.e., referential use of space). In addition, in each culture speakers used affirmative head movements to elicit backchannels from their listeners: Bulgarian speakers used lateral shakes, the other speakers used head nods.

Comparative studies, such as [Safadi and Valentine, 1990], suggest cross penetration of previously ethnically specific behaviors. In particular, the upward toss of the head as “no” may not be as prevalent as before. We hope our corpus analysis will give us more up-to-date intuition on this issue.

3.1.1.3 *Nonverbal behaviors in service encounters*

In general, gaze (see Montague et al. [2011] for an overview) and smile (see, for example, Kim [2009]) can play defining roles in the outcomes of service encounters. For instance, customers reported higher satisfaction when they interacted face-to-face with a bank teller who responded with contingent smile, rather than constant neutral or constant smiling expression [Kim, 2009]. The same data showed that amused and polite smiles differ with respect to their temporal proper-

ties [Hoque et al., 2011]. Analysis of verbal and nonverbal expressions in the videos of inter-ethnic encounters of Korean retailers with Korean and African-American customers showed that these language communities had different perception of function of socially minimal and socially expanded encounters [Bailey, 1997].

The openly accessible CUBE-G corpus of nonverbal behaviors from role plays of German and Japanese participants covers scenarios that may be relevant for service encounters, including first meeting, negotiation and status difference [Rehm et al., 2009a]. However, we have not been able to locate any nonverbal corpora of service encounters between speakers of Arabic. Hence, we collected our own video corpus of receptionist encounters, which may be the first annotated corpus of nonverbal behaviors in receptionist interactions, and the first annotated nonverbal corpus of service encounters freely available online at [Makatchev et al., 2012] (original video and audio data are not included). We describe this corpus in detail below, in Section 3.2.

3.1.2 *Verbal behaviors*

A number of corpora of service encounters exists. While some of them are concerned with encounters between members of different languacultures, the work on Arabic service encounters is scarce. One example is audio recordings of Syrian shopping interactions were collected and analyzed by Traverso [2001]. Shopping situations, have some similarity to the receptionist encounters. For example, in the simplest case, they both consist of several routine actions, with ritualized language patterns. In the shopping situations in Syrian Arabic, it's opening sequence, welcoming, offer of service, acknowledgment of the request, agreement for buying, exchange of money, closing sequence. In the receptionist encounters in English, as we will see, the sequence of steps is less involved: opening sequence, offer of service, request for directions, the directions exchange with possible clarification questions, and closing exchange. Traverso discusses that Syrian shopping exchanges can deviate from the routine, breaking into mutual "challenging activities," separate from bargaining and independent of the customer's decision to buy, that start with a face threatening act (e.g. criticism or confrontation). There is no telling if such activities transfer into a second language. If we were to find this kind of languaculture-specific behavior in Arabic receptionist encounters, and if it was also present when native speakers of Arabic were speaking English as a second language, it would be welcomed as a rich point that is a viable candidate for an ethnic cue.

In general, the following verbal behaviors are reported by Feghali [1997] as shared by native speakers of Arabic: repetition, indirectness, elaborateness, and affectiveness. It is not clear to what extent these behaviors transfer into second language. The literature on pragmatic

transfer from Arabic to English suggests that some degree of transfer occurs for conventional expressions of thanking, apologizing and refusing (e.g. [Bardovi-Harlig et al., 2007] and [Ghawi, 1993]). Thus, Ghawi [1993] notes that native English speakers used explanation strategy (e.g. “I’ve been busy”) only in about 29% of the apology situations, Native speakers of Arabic used explanation in 71% of situations when speaking Arabic, and in 76% of situations when speaking in English as a second language. This data provides us with a backup intuition for the design of the behavioral ethnic cue candidates, which we will rely on, as most of the results of our corpus study are suggestive at best.

3.2 A CROSS-CULTURAL CORPUS OF ANNOTATED VERBAL AND NONVERBAL BEHAVIORS IN RECEPTIONIST ENCOUNTERS

To address the lack of nonverbal corpora of human receptionist interactions, we collected and annotated a corpus of dyadic interactions between participants who play roles of a receptionist and a visitor. Our hope was that, guided by intuitions described in the related work (Section 3.1), we would identify distinctive verbal and non-verbal behaviors that could serve as viable candidates for behavioral cues of ethnicity, that we could additionally evaluate for their ethnic salience in a followup study. As we show in our analysis, some of the expected differences, such as in the amount of gaze on user, were not found. In other cases, where the differences are apparent, the sparsity of the data precludes us from claiming they are ethnically specific. In the end, we had to rely on a combination of the intuitions from the related work and the suggestive findings from our corpus analysis to generate a broad range of behavior candidates, in hope that true ethnic cues will be identified in our study on perceived ethnicity.

3.2.1 *Data collection*

3.2.1.1 *Participants*

We recruited via emails and posters in Education City in Doha, Qatar and via announcements posted on bulletin boards across the CMU campus in Pittsburgh, USA. The recruitment materials specified that we were looking for native speakers of American English or Arabic who are at least 18 years old. The majority of the participants (17 of 22) were university students, staff, or faculty. The participants filled out demographic surveys (figure 47, Appendix C) and evaluated themselves on a ten-item personality inventory (TIPI) Gosling et al. [2003] and 20-item positive and negative affect scale (PANAS) Watson et al. [1988]. The distribution of participants is shown in Table 1.

Doha	Arabic	Females	2 (2)
		Males	6 (5)
	American English	Females	2 (2)
		Males	3 (1)
Pittsburgh	Arabic	Females	1 (1)
		Males	1 (0)
	American English	Females	5 (2)
		Males	1 (0)

Table 1: Distribution of participants between Doha and Pittsburgh experiment sites. The numbers of different participants in the receptionist role, whose interactions were annotated, are shown in parenthesis.

People apply different criteria when they report their native language and mother tongue [Laitin, 2000]. For example, people may refer to their ancestral language as their “mother tongue” even if they do not speak it [Dorian, 1981]. To control for this, in addition to asking their native language, we asked the participants to list the countries they lived in for more than a year, and their ages at the time of moving in and out of the country. All but 3 participants (who were all in the American English condition in Doha) spent the majority of their lives in the country where their native language is a primary spoken language. A female participant in Doha changed her reported native language from American English to Tulu, after asking the experimenter a clarification question. Her data remains in the corpus although she is not included in the Table 1. One of the recruited subjects was an Arabic-speaking male professional receptionist. A female native speaker of American English had prior experience working at a reception desk. We will refer to these two participants as experts. Interactions of several participants were not analyzed due to lack of time or intentionally excluding them. In particular, data corresponding to two male participants playing a receptionist were excluded from the analysis: one American English condition participant had a strong East Asian accent and another Arabic condition participant held himself too casually for a typical receptionist. However, interactions where these two male participants were acting as visitors are included in the analysis.

Mean age of participants in Doha was 25 years (sample standard deviation, $SD = 7.8$). In Pittsburgh, average age was 28.7 years ($SD = 12.7$). Native speakers of Arabic were on average 23.2 years old ($SD = 4.2$), while the average age of native speakers of American English was 30.9 years ($SD = 12.5$).

3.2.1.2 Procedure

After filling out the questionnaires, one of the participants was asked to play the role of a receptionist while the other was ushered away from the reception desk and instructed to ask the receptionist for directions to a certain location inside or outside the university building. The location was picked by the experimenter from the following list: library, restroom, cafeteria, student recreation room, a professor A's office, another building on campus. Visitors were asked to seek help of the receptionist for directions using English language and then to proceed towards their destination.

Most of the participant pairs were not familiar with one another. The fact of familiarity, when clear, is noted in the annotations. Similarly, the annotations include information on whether the participant had a thorough (works or studies inside the building) or passing (works or studies in a nearby building) familiarity with the experiment site.

The sessions were held at 3 sites:

- Site 1: CMU Qatar lobby, the seat of an actual receptionist. A chair was provided. There was a computer screen with the building directory standing on the counter, in front of and facing the receptionist.
- Site 2: CMU Qatar lobby, the counter facing the direction opposite of Site 1, the security guard chair was vacated for the participant acting as a receptionist. There was no computer, but there was a magazine placed on the counter in front of the receptionist. This site was used as an alternative to Site 1, to reduce the disturbance to normal operations.
- Site 3: CMU Pittsburgh, an unoccupied receptionist counter on the 5th floor of Gates-Hillman Complex. There was no computer, but there was a magazine placed on the counter in front of the receptionist. A chair was provided.

In Doha, on-duty security guards were present in the vicinity of the reception desk.

Each pair of participants would have 2-3 interactions with one of the subjects as a receptionist, and then they would switch roles and have 2 or 3 more interactions, depending on allotted time. After that, the participants were debriefed on their experiences. Overall, more than 60 interactions were recorded.

The interactions were recorded with 2 or 3 consumer-level high definition video cameras. Visitor and receptionist were each dedicated a camera capturing their torso, arms and face that was positioned about 45 degrees off their default line of sight (namely, the line of sight that is perpendicular to the front edge of the rectangular reception desk). Most of the interactions would have a third camera capturing the side

view of the scene. All cameras were in plain view. In addition to the audio captured by the cameras, an audio recorder (iPod) was placed on the receptionist desk.

3.2.2 Annotation scheme

Our main motivation for this corpus collection is to mine occurrences and time profiles of verbal and nonverbal behaviors for potential rich points. Consequently, we have chosen to annotate the data at the level of granularity that minimizes the coding effort while at the same time allowing to capture timing and major features of communicative events. For example, instead of annotating each of preparation, hold, stroke, and retraction phases of a hand gesture (see [Kita et al. \[1998\]](#) for the description of phases of body movements), we annotated an interval between the beginnings of the stroke and retraction phases. Similarly, facial expressions were annotated as intervals approximately from the beginning of rise to the beginning of decay phases, with some error inherent to manual annotation (see [Hoque et al. \[2011\]](#) for the description of phases of facial expressions). The annotation scheme, developed in the process of annotating the corpus, is summarized in Table 2.

Modality	Values
Speech	Transcribed utterances, including non-words
Eye gaze	Pointing (self-initiated), pointing (following interlocutor), focus (interlocutor, guard, desktop, down, up, left, right, front, back, scattered, destination)
Face	smile (open or closed lips)
Head	nod, half nod, double nod, multiple nod, upward nod, multiple upward nod, micro nod, shake
Hand	Pointing (left or right hand), finger only
Torso	Sitting, standing, focus (left, right, front, back, destination, interlocutor, desk)

Table 2: Annotation scheme.

Coding nonverbal expressions, as well as transcribing ambiguous speech, involves a degree of subjectivity. For example, the exact point of gaze fixation within the recipient's face is hard to identify even by the recipient himself (as shown in the study by von Cranach and Ellgring [1973]). In fact, Cook [1977] showed that a typical direct eye contact consists of a sequence of fixations on different points on the face. Since it is unclear whether the exact fixation pattern has any influence on social communication, in this study we do not distinguish between different fixation points within the general face area (neither does the video fidelity allow that). The author was the primary annotator. Three of the interactions were annotated by a second annotator (associated with the study, but unaware of the exact study hypotheses discussed in Section 3.1.1) with the F-scores between the two annotators' temporal interval annotations for the receptionist's speech, gaze on user and pointing gazes equal to 0.94, 0.96, and 0.88, respectively. The F-score of the transcribed receptionists' speech histograms is 0.90.

The annotation is done using the multi-track video annotation tool Advane [Aubert and Prié, 2005]. The tool allowed us to simultaneously annotate multiple modalities in a timeline interface. It was chosen over other annotation tools with similar capability due to its free availability on Linux and relative stability of the code.

It rarely crashed.

3.2.3 Note on the analysis

Analysis of such rich multimodal data is a complex task. We make it feasible by limiting our analysis below to the modalities that are afforded by our robot prototype (see Section 1.1.1), namely gaze, neck movements, facial expressions and verbal behaviors. The variety of sites and destinations, combined with a relatively low number of subjects, requires care when comparing the data, since even the same destinations will generally require different directions from different sites. Therefore, we control for the total duration of the interaction for aggregate durations of events, and focus largely on events of a smaller scale, such as fragments of continuous gaze, occurrences of facial expressions and nods, and relative word frequencies. We conclude with a qualitative analysis of a few cases of greetings, disagreements, and failures to provide information.

3.2.4 Analysis: duration of interaction

The 46 annotated interactions with 13 participants acting as receptionists lasted about 1697 seconds (28.3 minutes) in total. Both Arabic and American English-speaking female receptionists had about the same average duration of interactions (figure 2). Three interactions with the AmE male receptionist notwithstanding, on average, female receptionists tended to interact longer than male ones (the

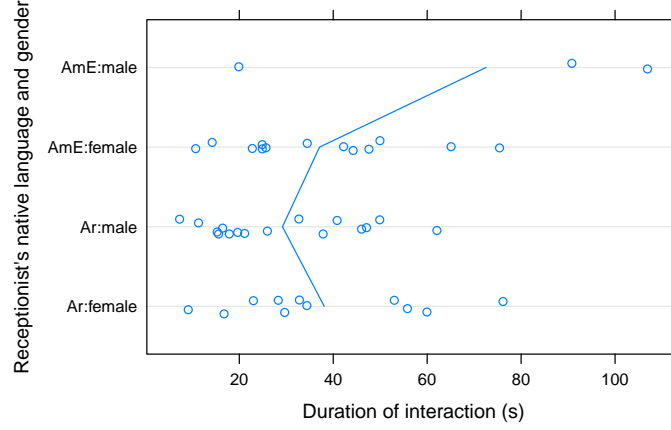


Figure 2: Duration of interactions with receptionists grouped by native language and gender. Interactions where receptionists were not able to give any directions are excluded.

95% highest posterior density (HPD) interval for the male coefficient is $[-37.31, -2.49]$ and the linear model (LM) coefficient's p-value is 0.026). Interactions were longer with visitors who are native speakers of American English (the 95% HPD interval is $[3.30, 29.26]$ and the LM coefficient's p-value is 0.063). Significant interaction were found between male visitors and receptionists, with a positive effect on duration (the 95% HPD interval is $[11.20, 52.44]$ and the LM coefficient's p-value is less than 0.001).

3.2.5 Analysis: gaze

3.2.5.1 Overview of gaze behaviors

Two gaze behaviors that make up the longest fraction of interactions are *gaze on visitor* and *gaze towards the landmark while pointing (pointing gaze)*. Other gaze behaviors annotated included gaze at the desktop computer (for Site 1) and gaze towards a particular direction that is not the user or while pointing (referred to as *other gaze*). We briefly review these gaze behaviors here and focus on each of them in more detail in the sections that follow.

Fitting a linear model that predicts the amount of a receptionist's gaze on visitor given the duration of interaction (the scatter plot in figure 3) yields the significance of the duration ($b = 0.79$, $t(44) = 14.29$, $p < 0.01$). The overall model's fit also shows significance with adjusted $R^2 = 0.8$, $F(1, 44) = 200.2$, $p < 0.01$.

As seen in figure 3, the biggest outliers from the fitted line are receptionists r1, r2, r4. All of them were in Site 1, and so either used the computer (the expert Ar male receptionist r1) or looked at the

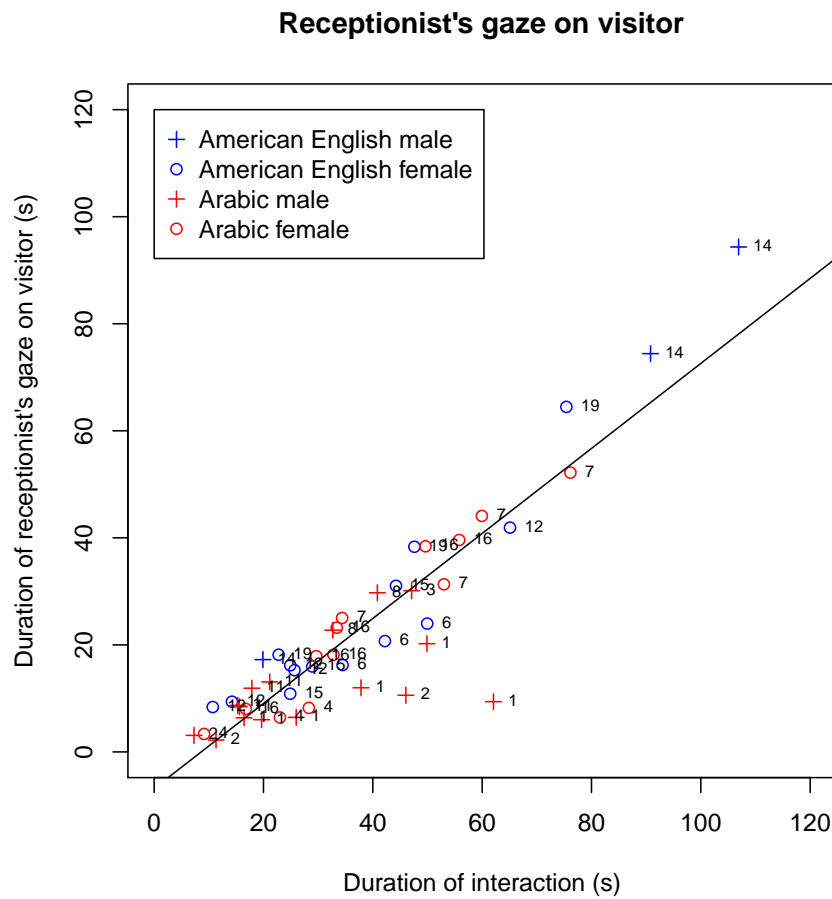
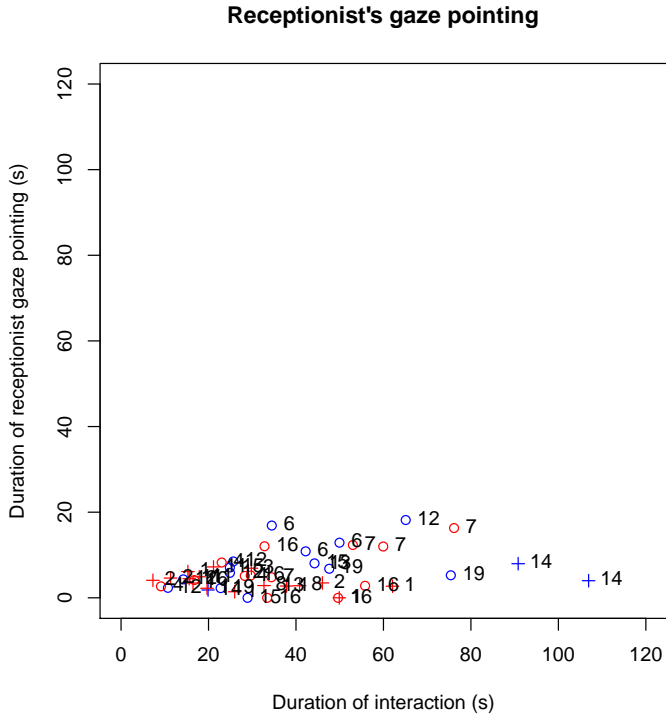
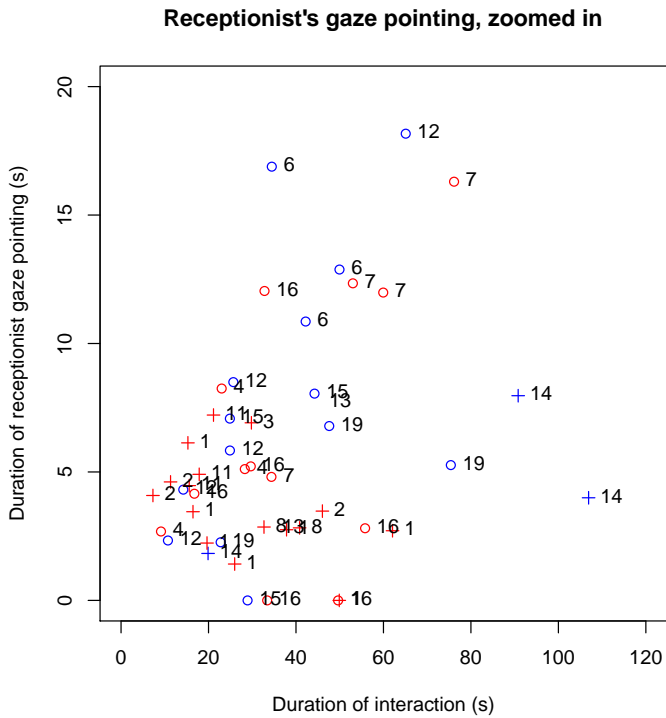


Figure 3: Duration of a receptionist's gaze on the visitor versus the length of the interaction. Numbers correspond to the participants' ids. The line corresponds to the best linear fit, $R^2 = 0.8$, $F(1,44) = 200.2$, $p < 0.01$.



(a) Duration of a receptionist's gaze pointing versus the length of the interaction (equal scale axes).



(b) Duration of a receptionist's gaze pointing versus the length of the interaction (zoomed vertical axis).

Figure 4: Gaze on visitor and pointing gaze.

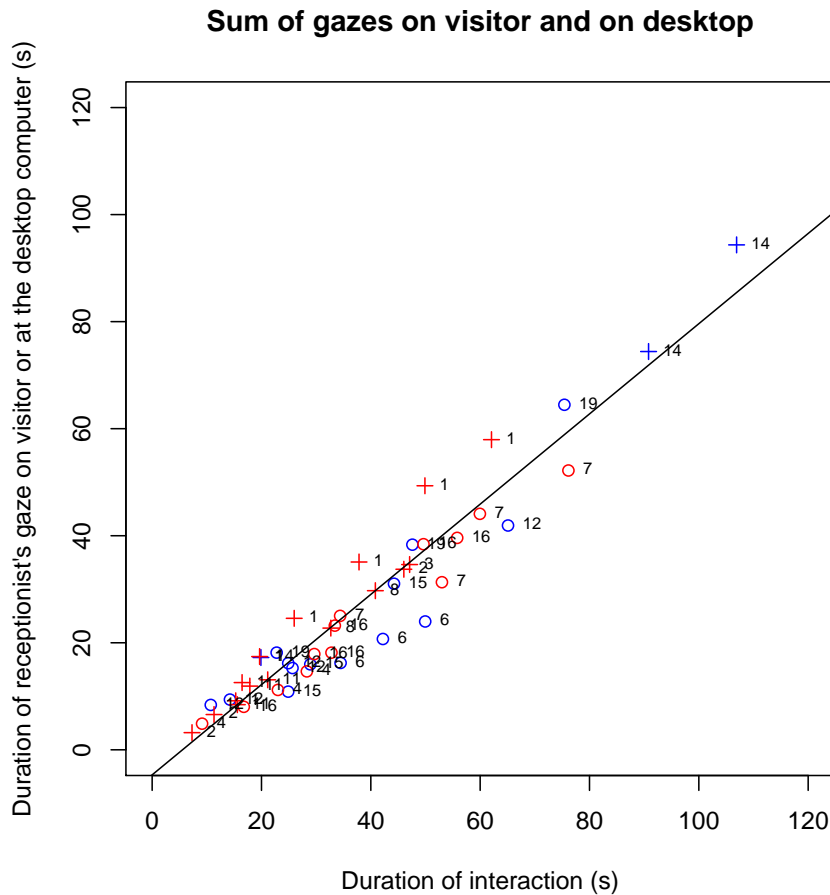


Figure 5: Duration of a receptionist's gaze on the visitor plus on the desktop versus the length of the interaction. The line corresponds to the best linear fit, $R^2 = 0.92$, $F(1,44) = 535.9$, $p < 0.01$.

screen (receptionists r2 and r4). The cumulative durations of on user and on the desktop computer produces a linear fit shown in figure 5. The duration is significant with $b = 0.85$, $t(44) = 23.15$, $p < 0.01$, and the complete linear model is significant with $R^2 = 0.92$, $F(1,44) = 535.9$, $p < 0.01$.

It is interesting to observe in figures 4a and 4b that the duration of pointing gaze does not grow proportionally with the duration of interaction, and there is a relatively large range of pointing durations. All longer pointing durations (8 seconds or longer) correspond to female receptionists. More analysis and data would be necessary to control for destinations, since they affect complexity of directions.

3.2.5.2 Other gaze

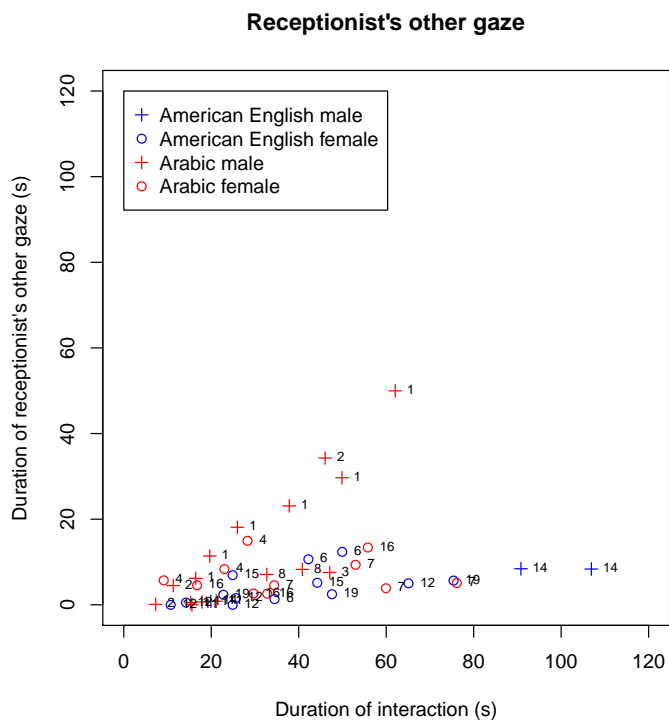
Other gaze corresponds to gaze anywhere but towards the visitor (*gaze on visitor*), and not towards the destination while giving directions (*gaze pointing*) or while listening to the visitor repeating the directions (*gaze towards where visitor is pointing*).

Figure 6a shows that apart from the clear outliers corresponding to Site 1 (receptionists r1, r2, r3, and r4), durations of other gazes increase relatively slowly with the duration of interaction. However, observe that even the same receptionists exhibit a broad range of their *other gaze* duration (e.g. r6 and r16).

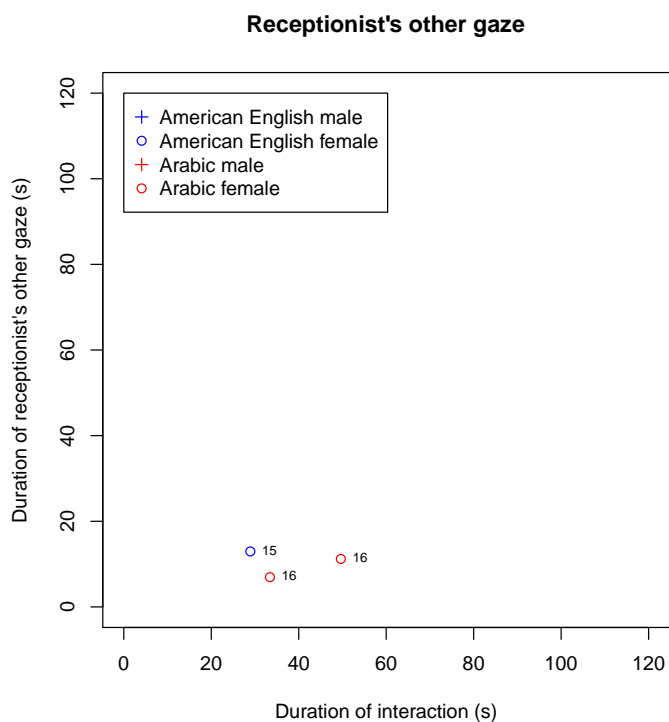
Figure 6b shows durations of *other gaze* for the 3 interactions where no directions were given as the receptionists admitted not knowing the information. Notice the trend towards more of *other gaze* when the receptionist did not know the directions. Thus, r15 spent 13.0 seconds performing *other gaze* when she did not know the directions, versus 5.2 and 6.9 seconds when she did. Similarly, r16's *other gaze* took 7.0 and 11.2 seconds when she did not know directions, versus 13.4, 2.7, 4.6 and 2.6 seconds when she did.

The data showed no associations of the amount of *other gaze* with gender or ethnicity of either receptionists or visitors.

Figure 7 shows the estimated density of distribution of durations of continuous other gaze intervals for Arabic and AmE female receptionists when they were able to give directions. Both distributions peak at about the same value within 0.5–1 second range, although AmE females exhibit a noticeable density peak in 2–3 second range.



(a) Interactions where some directions were given.



(b) Interactions where receptionists failed to give directions.

Figure 6: Duration of a receptionist other gaze (including gaze on the desk-top) versus the length of the interaction.

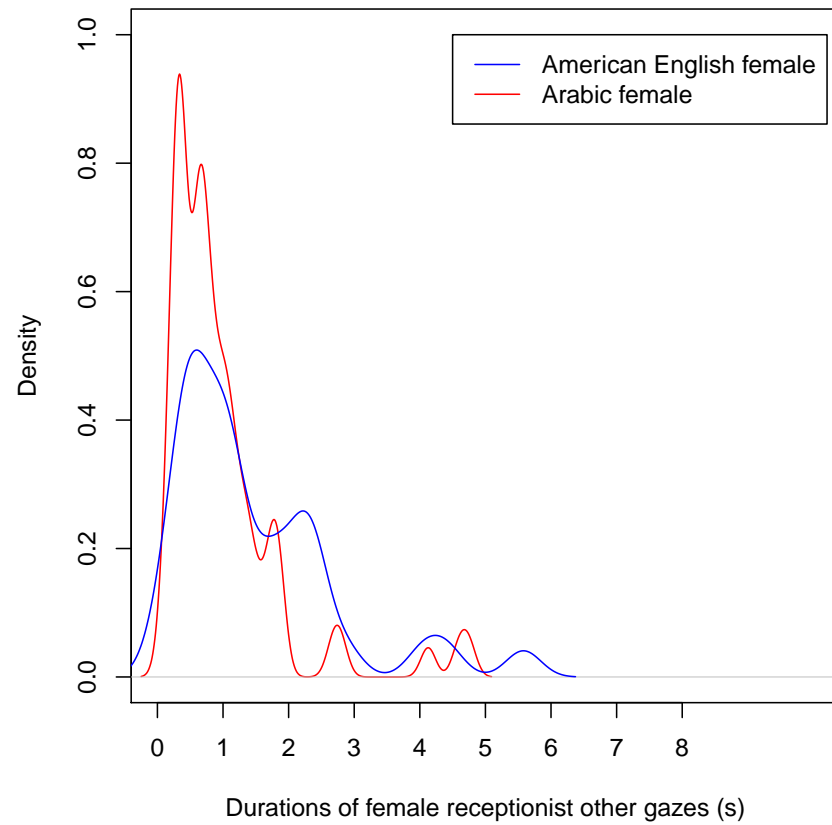


Figure 7: Gaussian kernel density estimate of female receptionist's continuous other gaze durations. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically.

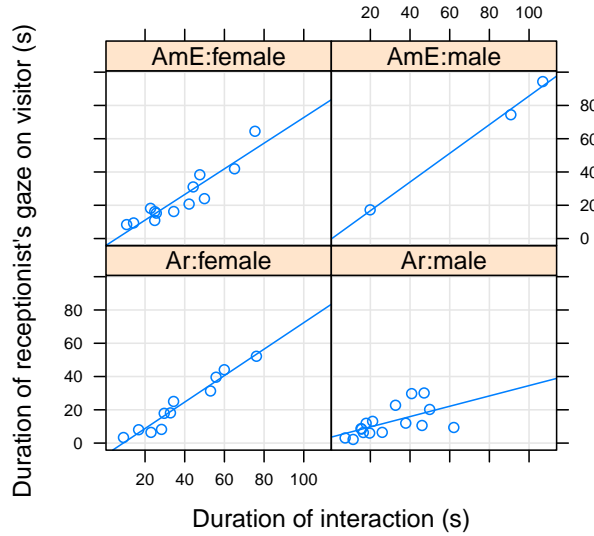


Figure 8: Duration of a receptionist's gaze on visitor versus the length of the interaction. Interactions where receptionists were not able to give any directions are excluded. The data is grouped by the receptionist's native language and gender.

3.2.5.3 Gaze on visitor

The amount of gaze on user does not reveal any association with native language or gender of the receptionist or the visitor. The gentler slope of the linear fit for male native speakers of Arabic (figure 8) is in part due to the data points corresponding to Site 1, where all the receptionists gazed on the desktop computer. An optimistic reader could draw connections with our hypothesis that Ar receptionists would exhibit less amount of gaze on visitor. We caution against such wishful thinking, as the 3 interactions for AmE male receptionist are based on a single subject and the difference is not present among females.

As we discussed in Section 3.1.1.1, amount of gaze on listeners during speaking can be associated with status differences. Therefore, given the reported differences in power distance between AmE and Ar languacultures [Hofstede, 2001], we expected to see an effect of ethnicity on the amount of the receptionist's gaze on visitors. However, this was not the case. Excluding the cluster of receptionists r1-r4, who spoke while looking at the computer, and therefore gazed less at the visitor, there is no significant associations between ethnicities (either the receptionist's or visitor's) and the amount of gaze on visitor while speaking (figure 9). There is also no significant associations with either receptionist's or visitor's genders.

Analysis of durations of continuous gazes on visitor shows that Arabic speakers tend to have more of short-duration gazes lasting less

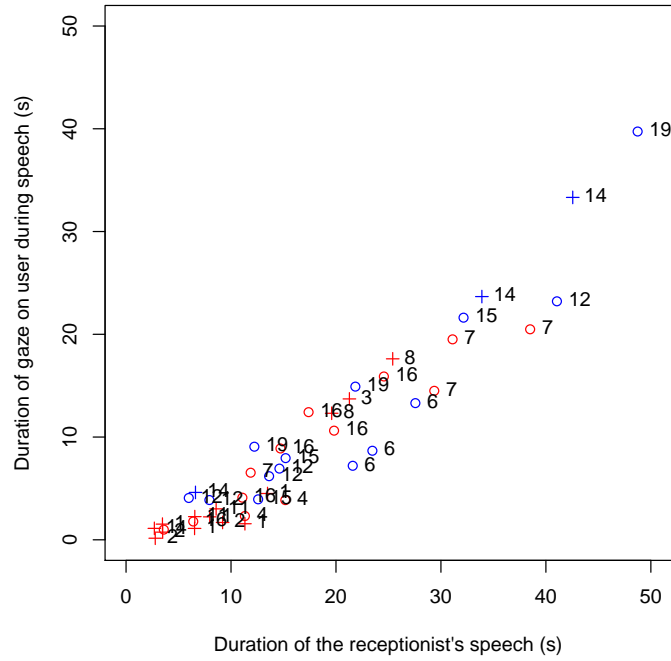


Figure 9: Duration of a receptionist's gaze on the visitor during the receptionist's speech.

than 1 s (figure 10). This is true for both female and male subjects (figures 11 and 12). In fact, the mean duration of gazes on visitor for Arabic speakers is 2.78 s versus 4.40 s for speakers of American English ($p < 0.002$). However, the difference of means for female speakers is not significant (figure 13).

Gender of the visitor does not appear to affect the distribution of continuous gaze on visitor.

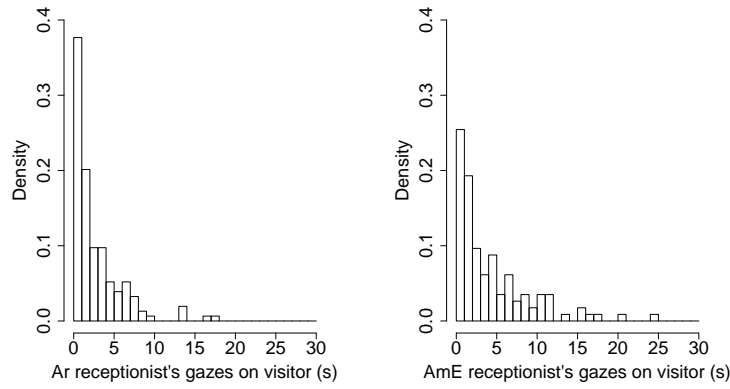


Figure 10: Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English.

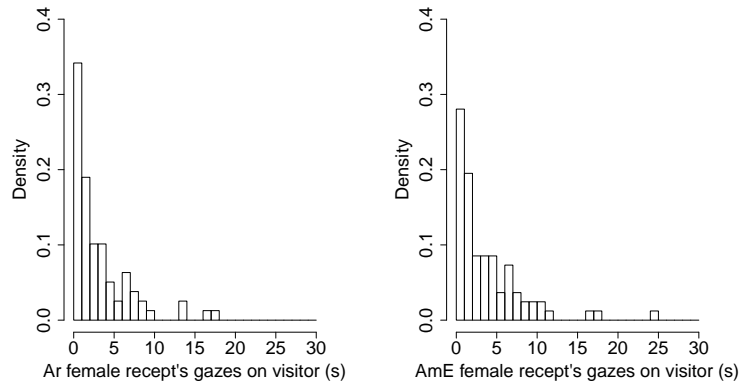


Figure 11: Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English. Female receptionists only.

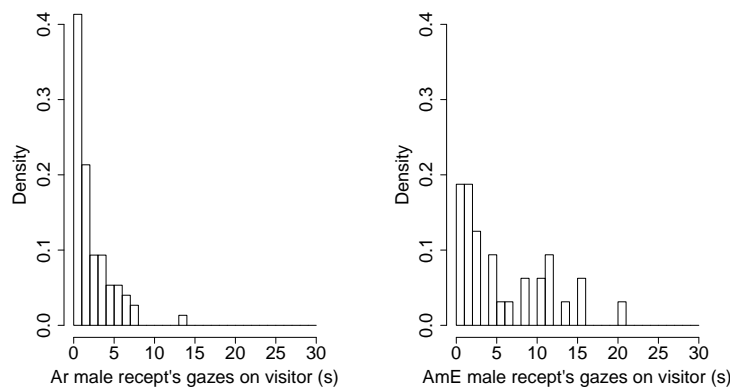


Figure 12: Distribution of durations of continuous gaze on visitor for speakers of Arabic and American English. Male receptionists only.

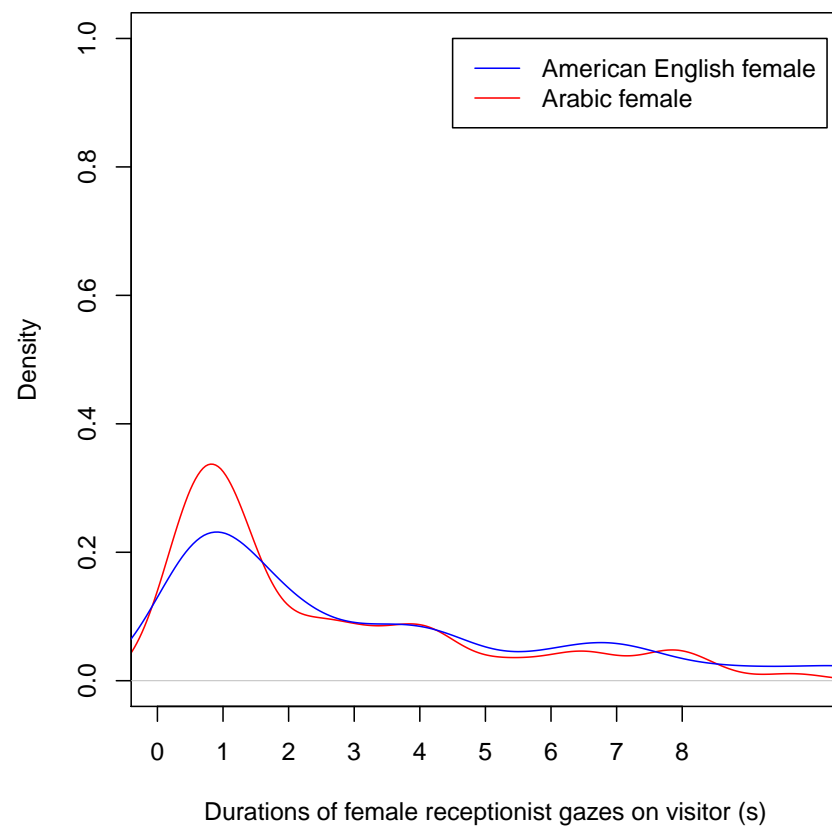


Figure 13: Estimated distribution of durations of continuous gaze on visitor for female speakers of Arabic and American English. Only interactions where directions were given are shown.

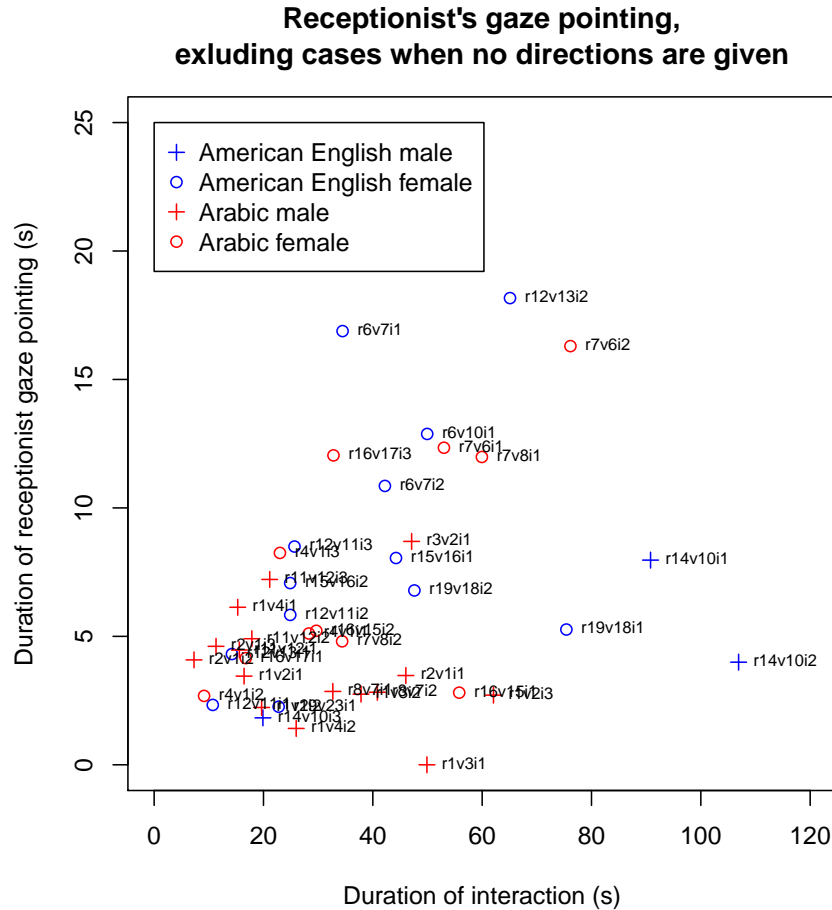


Figure 14: Duration of a receptionist gaze pointing versus the length of the interaction. As opposed to figure 4b, interactions where receptionists were not able to give any directions are excluded. Labels correspond to the full interaction code: the receptionist's id, the visitor's id and the id of the encounter for a given pairing of receptionist and visitor.

3.2.5.4 Pointing gaze

Zero pointing for interactions r16v15i3 (the code stands for receptionist 16, visitor 15, and interaction 3), r16v17i2, and r15v16i3 correspond to the cases where receptionists did not know the directions. Figure 14 shows the aggregate duration of pointing gaze and does not include interactions where receptionist was not able to give any directions. Differences within an individual behavior can be seen when comparing r16v15i1, which had little pointing, with r16v15i2 and r16v17i3. In r16v17i3, the receptionist is confidently giving directions to the destination she has just learned herself (when she was playing a visitor).

The plots and linear fits (see Figure 15) suggest that there may not be an effect of language but there may be an effect of the reception-

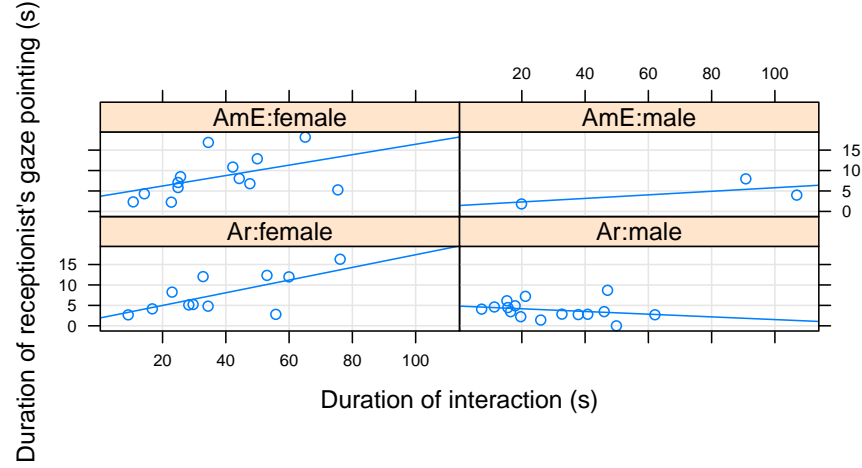


Figure 15: Duration of a receptionist's gaze pointing versus the length of the interaction. Interactions where receptionists were not able to give any directions are excluded. The data is grouped by native language and gender.

ist's gender. Indeed, fitting a mixed-effect generalized linear model shows that male receptionists pointed with their gaze significantly less than female ones (the 95% HPD interval for mixed-effects model coefficient is $[-12.13, -2.29]$ and the LM coefficient's p-value is 0.002). Also, receptionists of both genders tend to point less to male visitors (the 95% HPD interval for the mixed-effects model coefficient is $[-6.91, -0.09]$ and the LM coefficient's p-value is 0.044).

Analysis of durations of pointing gazes shows a trend towards short gazes (less than 1 s) for Arabic speakers, compared to the prevalence of gazes lasting between 1 and 2 seconds for speakers of American English (Figure 16). This trend is present for both female and male subjects (Figures 17 and 18, respectively). Correspondingly, Arabic speakers tend towards shorter (less than 1 s) gaps between gazes on visitor. Figure 19a demonstrates this tendency for Arabic females.

The skew of Arabic female receptionist's continuous pointing gaze durations towards sub-second glances is evident in the Gaussian kernel density estimate with an automated bandwidth selection, shown in Figure 19b. Velichkovsky et al. [1997] (as quoted by Pfeiffer [2010]) described associations between cognitive levels of information processing and durations of gaze fixations. However, he did not distinguish among fixations lasting longer than about 500ms, attributing all such gazes to communication tasks. While this discrepancy appears to be associated with ethnicity, its underlying mechanisms are unclear. Thompson [2012] suggests that higher actual or perceived

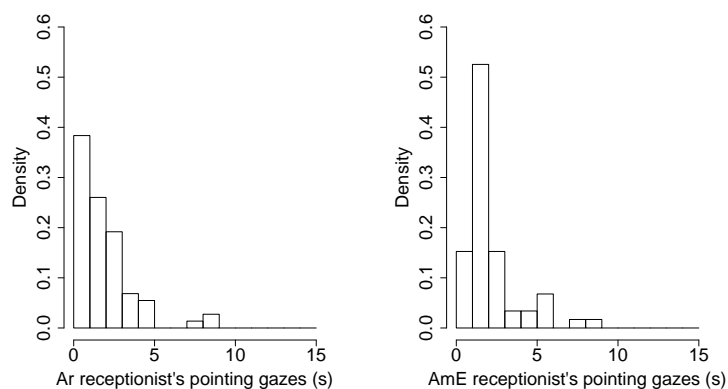


Figure 16: Distribution of durations of continuous pointing gaze for speakers of Arabic and American English.

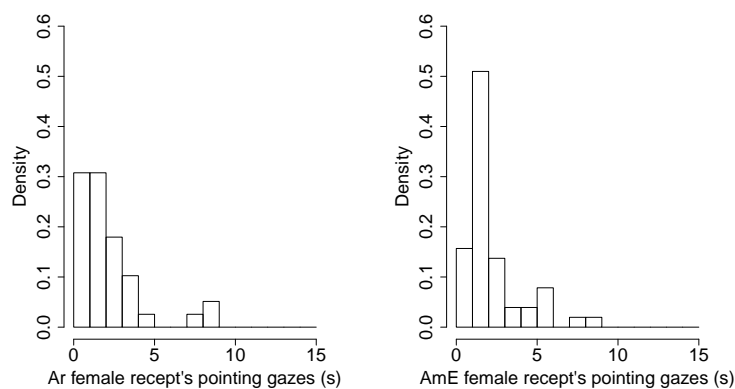


Figure 17: Distribution of durations of continuous pointing gaze for speakers of Arabic and American English. Female receptionists only.

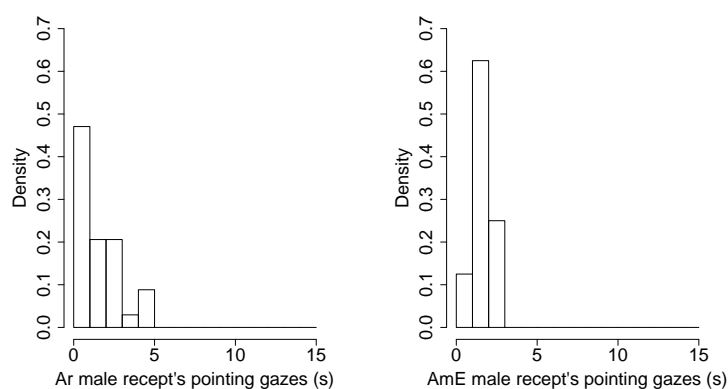
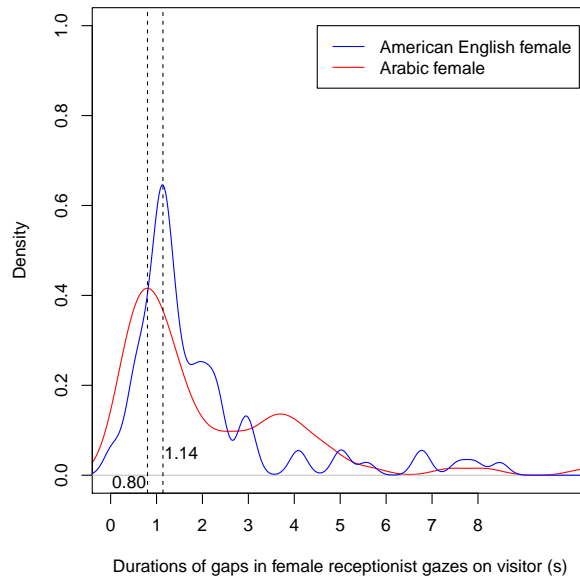


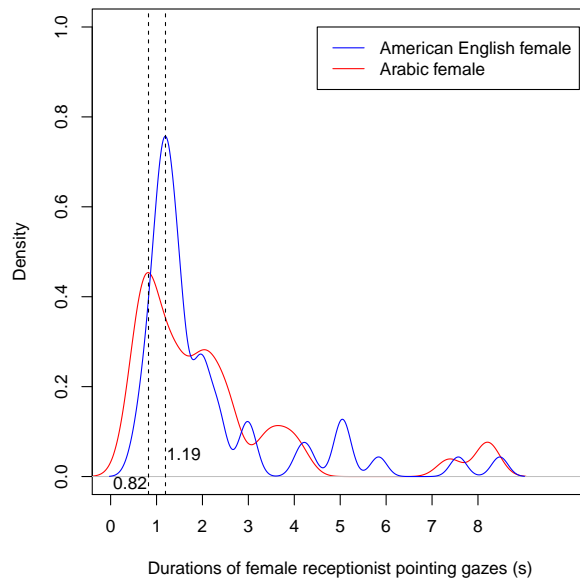
Figure 18: Distribution of durations of continuous pointing gaze for speakers of Arabic and American English. Male receptionists only.

speed of gaze transitions, that might result from the shorter gaps in gaze on listener, is one of the technics used in an actor's portrayal of lower status.

Gender of the visitor may have an effect on the distribution of the continuous pointing gazes. For example, Arabic female receptionists, when talking with male visitors, spend a fraction of time pointing over durations of 7 to 9 seconds, while with the female visitors the distribution of pointing gazes is concentrated in the interval between 0.5 and 3 seconds (Figure 20a). This pattern is not as pronounced in American female receptionists (Figure 20b). However, small amount of data per condition prevents us from making inferences.

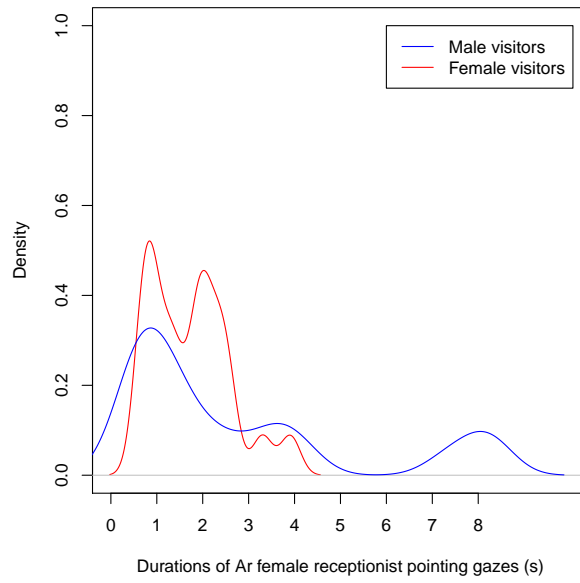


(a) Durations of gaps in gaze on user.

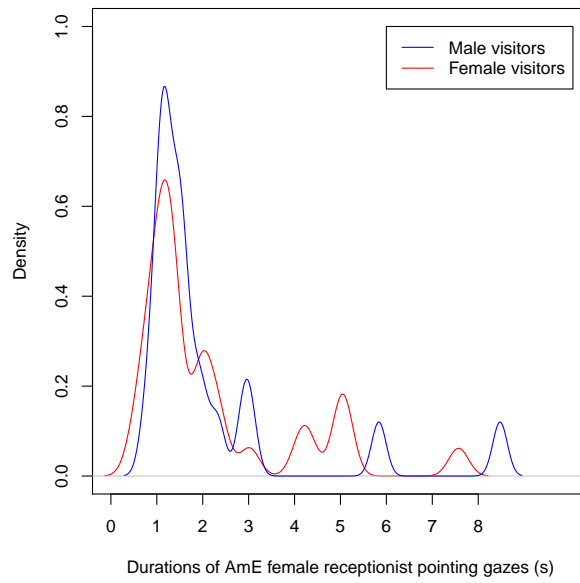


(b) Durations of continuous pointing gaze.

Figure 19: Gaussian kernel density estimates of female receptionist's durations of gaps in gaze on user and durations of continuous pointing gaze. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically. The dashed lines indicate the durations corresponding to the modes of the density estimates.



(a) Durations Ar female receptionist pointing gazes with male and female visitors.



(b) Durations AmE female receptionist pointing gazes with male and female visitors.

Figure 20: Gaussian kernel density estimates of female receptionist's continuous pointing gaze durations with male and female visitors. Interactions where receptionists were not able to give any directions are excluded. The bandwidth is selected automatically.

3.2.5.5 *Gaze in a conversation*

Case studies of gaze behaviors in the context of a conversation can yield some hypotheses for further evaluation. For example, compare the gaze behaviors of the Ar female receptionist r4 with an Ar male visitor v1 (figure 21) and AmE female receptionist r12 with Ar male visitor v11 (figure 22). Notice the short glances towards the visitor that punctuate fragments of the directions sequence spoken by r12. These glances appear to precede visitor's backchannels and therefore may play a role in maintaining engagement, i.e. are likely *connection events* Rich et al. [2010]. Receptionist r4, on the contrary, did not glance at the visitor until the very end of the directions sequence. These different gaze behaviors may reflect individual styles, genders and cultures of receptionist-visitor pairs, or levels of comfort and expertise, among other possibilities. While the small number of subjects prevents us from drawing such conclusions from our corpus, we will evaluate how individuals perceive duration and frequency of mutual gaze via a crowdsourcing study (Chapter 4).

3.2.5.6 *Summary of gaze analysis*

Contradicting to our hypothesis, there were no ethnic differences in the amount of the receptionist's gaze on visitor. Females tended to point more for longer interactions, while males used a constant amount of pointing gaze. Continuous pointing gazes tended to be shorter for Arabic receptionists: the mode for Ar females was 0.82s while the mode for AmE females was 1.19. We will use the duration and frequency of the continuous pointing (and, as a side effect, mutual) gaze as one of the variability dimensions in the study of ethnic attribution described in Chapter 4. Although the total amount of gaze on visitor did not find convincing support in our data, we will use it as another dimension of variability, since it has some support in the related work.

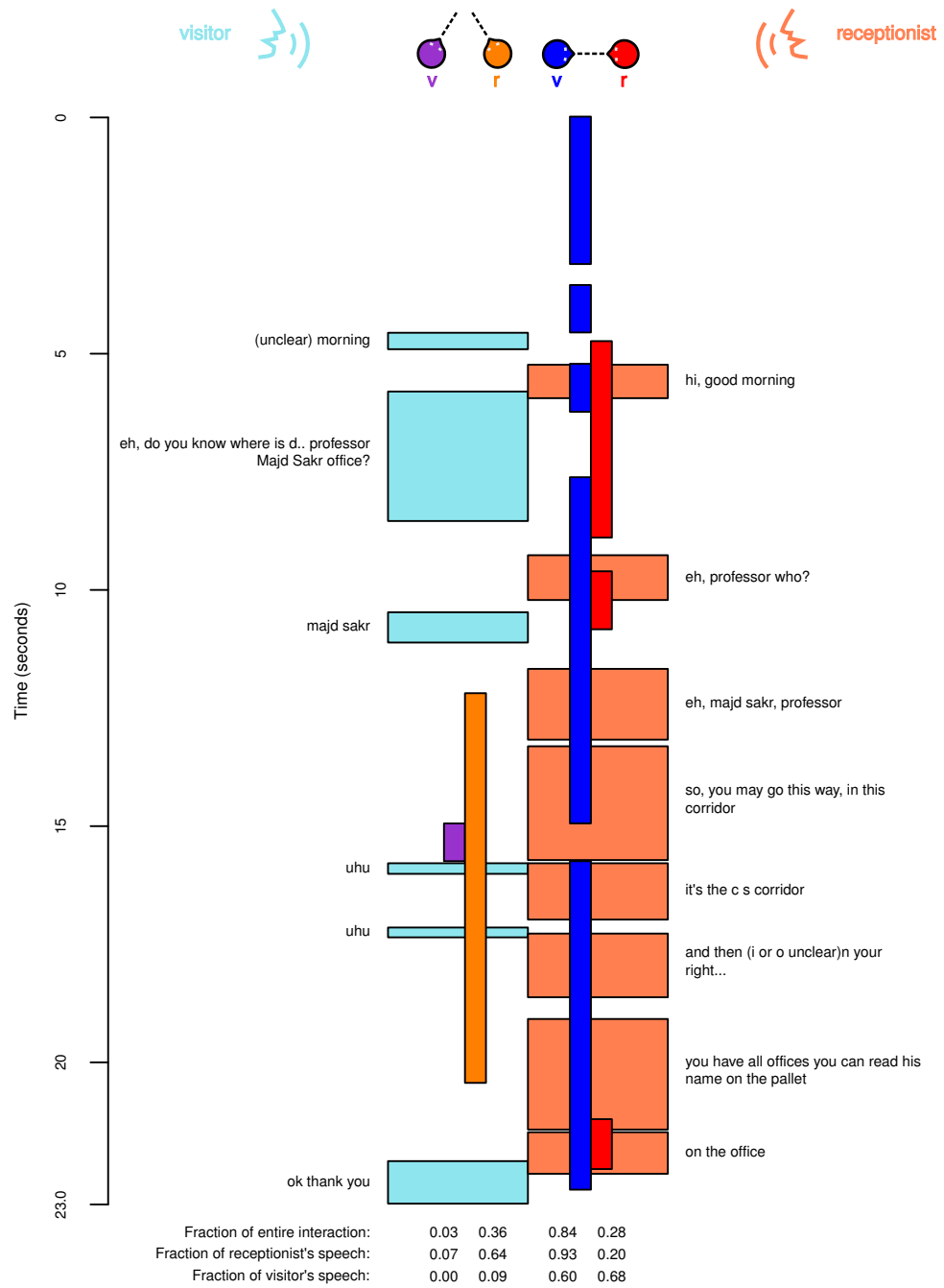


Figure 21: Interaction between the visitor v1 and the receptionist r4. Wide vertical stripes represent intervals of speech. Narrow vertical stripes represent (from left to right): intervals of visitor's and receptionist's gaze towards the direction pointed by the receptionist, and visitor's and receptionist's gaze towards each other. Color coding of these modalities is specified by the icons in the upper part of the plots.

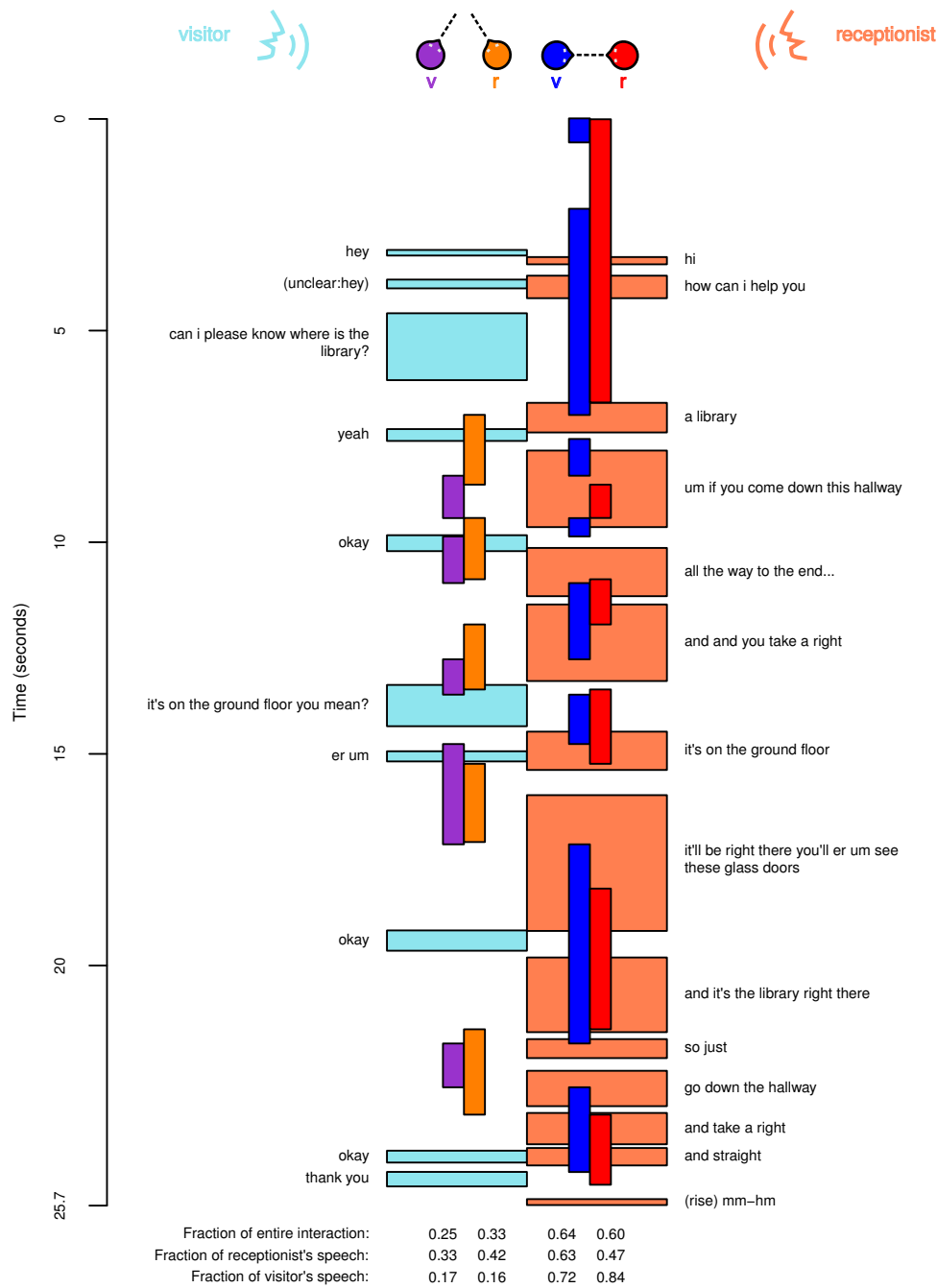


Figure 22: Interaction between the visitor v11 and the receptionist r12. The visitor's eye gaze for this particular dialogue is partially inferred from his head gaze. Wide vertical stripes represent intervals of speech. Narrow vertical stripes represent (from left to right): intervals of visitor's and receptionist's gaze towards the direction pointed by the receptionist, and visitor's and receptionist's gaze towards each other. Color coding of these modalities is specified by the icons in the upper part of the plots.

3.2.6 *Smiles and nods*

For smiles and nods, we focus on the expert receptionists: an Ar male and AmE female. The Ar male receptionist consistently combined his verbal greeting (typically with “yes, sir”, or reciprocatory “hi”) with a nod and slight closed smile. He also nodded as he backchanneled at the visitors as they were confirming the directions, and as a response to thanks during the closing (“you are welcome,” or “welcome, sir”). The AmE female, on the other hand consistently greeted with an open smile and “hi, how can I help you.” She only nodded during the direction exchange and the closing (“you are welcome”).

Contradicting to the expectations, we did not observe any upper head tosses, that may have the semantics of “no.” Some nods do appear to start with an upward motion, but we did not link it with any ethnicity.

3.2.7 *Lexical analysis*

Sometime during my life toilet paper became bathroom tissue.

— George Carlin

Potential differences in the dialects of English spoken by Ar and AmE participants may find their reflection in their lexical and syntactic choices. Similarly, possible pragmatic transfer (see Section 3.1.2), and differences in power distance are likely to affect word choices. We attempted to identify potential differences in the surface realization by performing N-gram frequency analysis.

- Only the Ar male expert receptionist used “sir.” The polite address was used in the greeting (“yes, sir”), direction giving (“sir please go to one zero zero seven”), and closing (“welcome, sir”). However, no receptionist said “madam,” or “ma’am.”
- While expressions “take a left (right)” were used by both non-expert AmE and Ar speakers, expressions “turn left (right)” and “go to your left (right)” were used exclusively by AmE speakers. The expert Ar speaker never used the “take a left (right)” expression, instead saying “go to the right,” “walk that way.”
- “You may go” and “you may take the (central elevator)” were used only by the Ar male expert receptionist and the Ar female receptionist who was paired with the expert before as a visitor (and who heard him use “you may take the (central elevator)”).
- “Go straight ahead” was used exclusively by AmE speakers, while “go straight” without following modifier “ahead” was used exclusively by Ar speakers.

- Only the Ar male expert receptionist (r1) used “please”:

r1: sir please go to one zero zero seven

r1: hold on please

r1: just a second please

In the last two cases, the receptionist paused while looking up the directions on the computer. It should be noted, however, that, with one exception, the computer was used only by the Ar expert receptionist.

- Three Ar receptionists responded to the closing “thanks” with “welcome,” or “welcome, sir.” No AmE receptionist used this response, normally saying the conventional “you are welcome.”

Using politeness markers “you may,” “please go,” as well as the deferential term of address “sir,” fits well within the predictions of Arab culture models that claim high power distance and importance of relative status. Indeed, [Brown and Levinson \[1987\]](#) claimed that the level of politeness is determined by “relative power ... the social distance ... and the ranking of the imposition.” Workplaces are no exception, with “power and politeness ... inextricably intertwined in every workplace interaction” [[Holmes and Stubbe, 2003](#)]. Finally, [Daller and Yildiz \[2006\]](#) show that Hofstede’s power distance values of nations can remain relevant in spite of global convergence.

Closing “welcome” instead of “you are welcome,” and phrasing “go straight” instead of “go straight ahead” while specific to Ar subjects in our corpus, at the time were interpreted by the author, perhaps naively, as indicators of English proficiency, rather than indicators of Arabic native language. However, “welcome” as response to thanks can be interpreted as a transfer from Arabic “ahlan,” that can be used both as a response to thanks and as a welcoming to one’s house. [Wertheim \[2013\]](#) referred to this phenomenon as *L1 semantic and pragmatic mapping into L2*. These phrases, together with “take a left (right),” “turn left (right),” and “go to your left (right),” would be good candidates for further evaluation of their ethnic salience via crowdsourcing, as described in Chapter 4. As we discuss in the end of this chapter, we considered the set of the candidate behaviors that we identified sufficiently large, and did not include every possible candidate into the crowdsourcing study.

3.2.8 Conversation analysis

Due to the scarcity of the dialogue data, we performed a conversation analysis of a number of representative cases. We focus our conversation analysis on dialogue acts that are common in receptionist interactions and are reported elsewhere as places of high cross-cultural vari-

ability (e.g. [Feghali, 1997; Bardovi-Harlig et al., 2007; Ghawi, 1993]): greetings, disagreements, and failures to provide information.

3.2.8.1 *Greetings*

Although Feghali suggested analysis of Arabic greetings, not much work on Arabic conversation openings, especially in service encounters, exists. A special case, phone conversation openings, has been compared between Palestinian and American English languacultures by Awadallah [2009]. Awadallah reports that Palestinian Arabic speakers, in contrast with speakers of American English, often duplicate “how are you,” namely, the Arabic analogue is said more than one time by the same speaker. Similarly Saadah [2009] reports that Arabic “how are you” exchange can take multiple turns. These differences in telephone conversation openings suggest that similar tendencies may exist in face-to-face interaction.

In our corpus data, Ar receptionists show tendency to greet with “hi, good morning (afternoon),” and “hi, how are you?” AmE receptionists tended to use simple “hi,” or “hey,” or “hi, how can I help you?” The expert AmE female receptionist used “hi” and a broad smile with lifted eyebrows. The Ar male expert receptionist greeted the same Ar male visitor with “yes, sir” and a nod in both of their encounters. In their first encounter, the receptionist initiated:

r1: Yes, sir
v3: Hello
r1: Hi

In their second encounter, the receptionist responded to the visitor’s greeting:

v3: Hello, I am so sorry
r1: Yes, sir

Responding to the visitor’s traditional greeting “as-salamu alaykum,” the expert Ar receptionist used the canonical “wa alaykum s-salam,” which can be translated as “peace be upon you,” and “and to you be peace,” respectively. No other instances of code-switching between English and Arabic were observed.

3.2.8.2 *Failure*

Failure to meet expectations is commonly followed by an apology, and apology strategies differ across languacultures. For example, Ghawi [1993] studied the use of following strategies: an expression of apology (“I am sorry”), an explanation (“I’ve been busy”), and acknowledgment of responsibility (“It was my fault”), an offer of repair (“Can I help you?”) and a promise of forbearance (“It won’t happen again”).

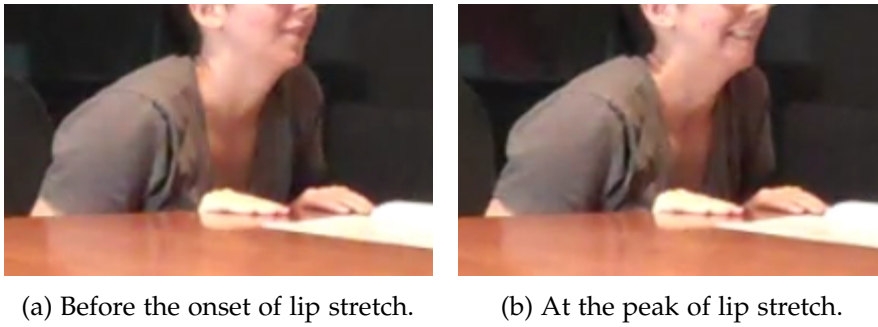


Figure 23: Receptionist r15 before and at the peak of lip stretch (AU20).
Used with permission.

Ghawi found an evidence of transfer of the frequently used explanation strategy into English spoken as a foreign language by native speakers of Arabic. We can expect that several of these strategies are applicable to the situation of handling the failure of giving the visitor the requested information.

On 7 occasions, the receptionist did not know some aspects of the directions to the destination. Some handled it by directing the visitor to a building directory poster or by advising to search for directions online. Others just admitted not knowing and did not suggest any workaround. Some receptionists displayed visible discomfort by not being able to give directions.

On two occasions receptionists offered an explanation for not knowing the directions. Thus, AmE female r6 said:

Explanation

r6: i am not hundred percent sure because i am new

Meanwhile, Ar female r16 said:

r16: i don't know actually his name, that's why i don't know where his office

An AmE female r15 used a lower lip stretch (figure 23) and an emphatic:

r15: I have n... I have absolutely no idea I've never heard of that professor and I don't know where you would find it.

She then went on to suggest to the visitor (v16) to look up the directions online. Interestingly, the visitor v16, who previously, as a receptionist, used an excuse and did not offer a repair, adopted r15's strategy and facial expression (lip stretch, AU20 of the Facial Action Coding System (FACS) by **Ekman and Friesen** [1978]) when acting as a receptionist again with the visitor v17:

Offer of repair

r16: i have no idea actually. you can look up like online or something

This example highlights the entrainment effects in our corpus. Since each subject participated in multiple interactions, usually with multiple interlocutors, the likely carry-over and entrainment effects may have reduced apparent intergroup differences. A deeper analysis, facilitated by a bigger amount of data, could account for the interaction history.

Although Ghawi [1993] found that native speakers of Arabic used an explanation strategy much more than native speakers of English, both when speaking their native Arabic and when speaking English as L2, the few failure cases in our data did not allow us to draw such inferences. The data is, however, useful for us to borrow the realizations of the apology strategies (phrasings) for the stimuli of the failure-handling behaviors used in our studies described in the following chapters.

3.2.8.3 *Disagreement*

On 7 occasions, a visitor tried to correct the receptionist, or incorrectly recited the directions when trying to confirm them. The receptionist responded in one of the three ways: (1) disagreed directly and followed with a correction, (2) disagree indirectly by providing the correct information, or (3) distanced himself from a disagreeable information by referencing the information source.

(1) The Ar female receptionist r7 directly disagreed and corrected Ar male visitor v8:

v8r7: Okay, so I just go all the way down there and...

r7v8: No, you take the elevator first.

(2) The expert Ar male receptionist r1, however, did not disagree directly with Ar male visitor v2:

r1v2: This way. Go straight.. (unclear) to the end.

v2r1: So this (unclear) to the left.

r1v2: Yeah, to the right. Straight, to the right.

Similarly, the AmE male receptionist r14 just gave the correct directions, when the visitor v10's clarification question had wrong information.

(3) The same expert Ar male receptionist distanced himself from the disagreeable information with another Ar male visitor¹:

v3r1: I think he moved upstairs, right?

Second floor. [He uses dean's office?]

r1v3: [Mm::] according
to my directory it's one one one zero zero seven.

¹ Square brackets are aligned vertically to show overlapping speech and colons denote an elongation of a sound.

We did not have clear predictions for differences in handling disagreements between Ar and AmE participants. On one hand, indirectness and power distance ascribed to Arabic languacultures could result in avoidance of face threatening forces that come from direct contradiction. Refusals, which are similar to disagreements in that they can also be realized via direct and indirect negation, are inappropriate in some situations for native speakers of Arabic. [Nelson et al. \[2002\]](#) note that such situations include refusing an interlocutor of higher status and refusing an invitation to dinner from a friend. In other contexts, however, [Nelson et al. \[2002\]](#) shows that both L1 English and L1 Arabic speakers use significantly more direct strategies than indirect ones.

Our data, too, shows that Ar receptionists can disagree directly and indirectly. We hope that further evaluation of the disagreement strategy would help to evaluate its usefulness for ethnic attribution.

3.3 SUMMARY

In this chapter, we reviewed the literature on rich points between Arabic and American English languacultures and collected our own corpus of receptionist interactions to identify rich points in this particular type of a service encounter. Some trends that correlate with languacultures were observed, in particular for distributions of durations of continuous pointing gazes. Other results are qualitative and are suggestive of associations between linguistic behaviors and such facets of the identity as expertise, gender, or ethnicity.

Analysis of our corpus data suggests that the difficulty of inferring rich points and their realization in a new context is inherent to both qualitative and quantitative studies. In short, no clearly distinct behaviors between Ar and AmE groups were found. However, our analysis and the related work allows us to formulate hypotheses on the realization of the potential rich point behaviors, which we will evaluate in the next chapter.

We hope that our multimodal annotated corpus of receptionist encounters will be useful to the community. Its annotations are freely available online [[Makatchev et al., 2012](#)].

EVALUATING ETHNIC SALIENCE OF RICH POINTS

In previous chapter we discussed that rich points and rich point candidates between AmE and Ar communities, identified in related work and in our corpus analysis, require further evaluation, for the following reasons: (a) rich points from related studies may not be applicable to our specific context of receptionist service encounters, (b) most of the behaviors identified via corpus analysis are only candidate rich points, as they may be associated with expertise or an individual, and finally (c) even true rich points may not be salient ethnic cues. The goal of the next stage of our methodology is to evaluate rich points and rich point candidates on their salience as cues of ethnicity. In particular, we would like to select behaviors that are cues of Ar ethnicity to members of Ar, AmE, or both, and, similarly, we would like to find cues of AmE ethnicity for Ar, AmE, or both ethnic groups. Such behaviors, when identified, can be implemented on the robot prototype which can be further evaluated in a smaller study with colocated participants.

Estimating ethnic salience of a large set of behaviors requires a large group of participants from both ethnic groups. In our work, we solve this problem by conducting the study online and recruiting the participants via an online workforce marketplace, Amazon’s Mechanical Turk (MTurk). This method of data collection is an example of *crowdsourcing*, defined by Howe [2006] as “the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.” Following crowdsourcing terminology we will sometimes address participants recruited via MTurk as *workers*.

Online crowdsourcing of our task behavior evaluation by ethnic communities is not as straightforward as we can hope. First, not every community is equally represented in online workforce, making the recruitment of workers from underrepresented communities harder and more costly. Second, online stimuli, such as text, or even video, may not have as much effect as a colocated robot would. The ability to select cues of ethnicity via crowdsourcing from rich points and rich point candidates is an integral part of the first hypothesis of this thesis:

Hypothesis I Believable ethnic identity of robot characters can be created using behaviors selected via lower fidelity on-screen simulations and crowdsourcing.

In this chapter, we describe our pilot study on applicability of crowdsourcing to evaluating personality and naturalness (used as a proxy for ethnic attribution) of dialogue transcripts presented as text. Encouraged by our results, we conducted a study that evaluated verbal and nonverbal candidate behavioral cues of ethnicity rendered on a robot and presented as videos. The stimuli in the latter study are evaluated on perceived animacy, anthropomorphism, likeability, intelligence, safety, and ethnic attribution. Results are not as easy to interpret for the purpose of selecting ethnic cues as we would hope. However, together with intuitions from related work on rich points, they allow us to narrow down the set of candidate ethnic cues and proceed towards implementing them in a robot prototype, and testing its ethnic attribution, as described in Chapter 5.

4.1 CROWDSOURCING HRI

As we defined in the beginning of this chapter, crowdsourcing refers to outsourcing tasks to a large group of people through an open call (as opposed to traditional employee recruitment). Crowdsourcing on the Web has recently gained popularity as a tool for conducting inexpensive and large-scale user studies. In human-robot interaction studies, Lee et al. [2010a], for example, used videos to introduce two robot prototypes, and described the actual test items using text and schematic imagery. The study evaluated user's perception of different strategies, realized as utterances spoken by the robot, for mitigating breakdowns in the task of bringing a can of soda. Chernova et al. [2010] used online multi-player games to collect situated dialogue data in the environment that mirrors an actual robot deployment scenario.

Crowdsourcing has been proven especially effective for labor-intensive natural language processing tasks, such as translation, information extraction and annotation [Callison-Burch and Dredze, 2010]. Crowdsourcing is particularly attractive for the opportunity to recruit study participants native to a particular locale. Higgins et al. [2010], for instance, reports that their tasks formulated in Arabic language attracted encouragingly high participation of workers from Arabic-speaking countries of Morocco, Egypt, Lebanon, Jordan, UAE and Dubai, in spite of the low overall proportion of workers from these countries [Ross et al., 2010].

Nevertheless, there are well known difficulties associated with crowdsourcing. First, special measures must be taken to ensure reliability of the data. The common undesirable issues for questionnaire tasks are multiple participation by one person, random answering, and demand characteristics (where subjects change their behavior in response to being measured) (see, for example, [Kittur et al., 2008]). We will outline how we addressed these issues in Section 4.2.5.

Attracting participants in an online market place like MTurk requires adjusting the monetary reward according to the workforce supply. In our studies we performed these adjustments manually, on batches of 10 or 20 tasks at a time. While reward for workers from the US was usually close to the minimum US wage (7.25 USD per hour), recruiting workers from Arabic speaking countries at times required up to 10 fold increase of pay. Other, non-monetary rewards, may be more successful in attracting scarce worker populations. For example, games with a purpose (see, for example, [von Ahn and Dabbish \[2008\]](#)) attract users by being fun to play.

4.2 EVALUATING VERBAL BEHAVIORS VIA TEXT STIMULI

4.2.1 *Motivation*

The goal of this pilot study is two-fold. First, it is to evaluate the feasibility of online crowdsourcing as a methodology for evaluating perception of linguistic stimuli across the communities of AmE and Ar workers. Second, shall the crowdsourcing work, the stimuli that receive different perceptual scores across user communities can be considered as rich points. Note that this study was conducted before the corpus collection described in Section 3.2. Therefore, we developed the stimuli for this study relying exclusively on related work on pragmatic transfer and cultural models (reviewed in Section 3.1). Hence, even though the study will allow us to identify linguistic stimuli that are perceived differently by AmE and Ar participants, we will not use the exact realizations of those stimuli in our further studies. Instead, we will use the rich points identified in this study as a guidance for the selection of categories of stimuli for a followup study, described in Section 4.3.

4.2.2 *Procedure*

After a demographic questionnaire (Appendix A, figure 47) each recruited Amazon's Mechanical Turk (MTurk) worker was presented with a sequence of web pages, one dialogue per page (Appendix A, figure 39). The instructions requested the worker to imagine that he is visiting a university building for the first time and sees a female receptionist who appears to be in her early 20s and of the same ethnic background as the participant. The following transcript is the dialogue that occurs between the participant and the receptionist. The dialogue is followed by the questionnaires: in one version of a study, it consists of the items of the Ten-Item Personality Inventory (TIPI [[Gosling et al., 2003](#)]), and in the other version it is a single item of naturalness on a 7-point Likert scale (for consistency with TIPI questionnaire).

The experiment used crossed design with the following factors: linguistic variability dimensions (verbosity, hedging, alignment, or formality—discussed in detail in Section 4.2.3), valence (negative, neutral control, or positive), dialogue acts (greeting, question-answer, disagreement, or apology), ethnicity (with levels American English and Arabic). Each participant is presented with 12 dialogues corresponding to 3 values of valence of one of the dimensions of variability across 4 dialogue acts. The naturalness version of the study, being less labor intensive with only one item in its questionnaire, presented each particular with 24 dialogues, corresponding to two dimensions of linguistic variability. Hence, valence and dialogue act were within-subject factors, while linguistic variability dimension were treated as an across-subject factor, as well as ethnicity and gender. Within each session the items were presented in a random order to minimize possible carryover effects.

In addition, each participant received one validation dialogue with a dialogue intended to be unambiguously unnatural if the participant indeed speaks American English or Arabic, depending on the ethnicity condition (see tables 11 and 12).

Upon completion of the study, the participant was presented with a random code to enter into MTurk's form in order to claim the payment. An experimenter had an opportunity to inspect the participant's data before authorizing the payment. A small number of participants failed validation questions and were banned from future participation.

4.2.3 *Stimuli*

This study was performed before the corpus collection and analysis described in Chapter 3, so the stimuli selected was guided purely by review of related work.

4.2.3.1 *Rationale*

We choose verbal stimuli according to their potential for being rich points between native speakers of American English and Arabic. [Feghali, 1997] summarizes the following linguistic features shared by native speakers of Arabic: indirectness, elaborateness, and affectiveness. The two questions relevant for our study are: (1) to what extent are these features transferred to native speakers of Arabic speaking English, and (2) what are the linguistic devices that realize these features.

The literature on pragmatic transfer from Arabic to English suggests that some degree of transfer occurs for conventional expressions of thanking, apologizing and refusing (e.g. Bardovi-Harlig et al. [2007] and Ghawi [1993]). We hypothesize that the transfer is present, or at least favorably perceived by native speakers of Arabic, in the dialogue acts relevant to receptionist dialogues. In addition to apologies,

these dialogue acts include greetings, answering information questions, and disagreement. Our experiment on scoring of naturalness of linguistic expressions is aimed at evaluating the perception of such transfer.

Indirectness, elaborateness, and affectiveness can be realized through multiple linguistic devices. Here, we are limited both by the size of the experiment with respect to the stimuli and the number of subjects, as well as by the behaviors that we can realistically generate in dialogue systems. One of devices that may be relatively easy to realize is verbosity. We use intra-turn verbosity for all dialogue acts except for greeting, for which, based on the ethnography by Suzanne Wertheim (reported in [Hobbs and Sagae \[2011\]](#)), we use discourse verbosity measured by the number of turns. Hedging has been found to be useful for expression of personality ([Mairesse and Walker, 2010](#)). We select it, speculating that it may be a plausible way of realizing indirectness in English by L1 Arabic speakers. Alignment (namely, using the same words or syntactic structures by multiple interlocutors), or its lack, are associated with affective language ([Isard et al. \[2006\]](#)).

Finally, cross-cultural differences in social distance and power status relationships affect the politeness strategies used by native speakers of English and Arabic (e.g. [Atawneh \[1991\]](#), cited in [Farahat \[2009\]](#)). Out of many linguistic devices that realize politeness strategies in English and Arabic, we have chosen a few that we united under the term *formality*. They include address forms (“sir” / “madam”), jargon and slang (“what’s up?”), and deference markers that target negative face [[Brown and Levinson, 1987](#)], such as “kindly (follow...)”, “I am afraid.”

An additional argument in favor of the dimensions of verbosity, hedging, alignment and formality as potential rich points is that they play a strong role in inferences about the personality of speaker (see, for example, an overview in [Mairesse and Walker \[2010\]](#)). Personality perception and personality types suitable for particular tasks are known often to be potential rich points [[Rushton, 1999](#)].

4.2.3.2 Realization

*And, boy, we went to talking.
She said “yeah,” and I said “yeah,”
And we went to having a real good time.*

— Blind Mississippi Morris and Brad Webb,
Juke, Back Porch Blues (1999)

We designed linguistic stimuli by varying sentence structure and lexical choice along the four dimensions of verbosity, hedging, align-

ment, and formality. We define these dimensions in more detail below.

- *Verbosity* corresponds to the number of words per turn (intra-turn verbosity) or number of turns (inter-turn verbosity). Positive valence of verbosity corresponds to more words (turns). We increased intra-turn verbosity by using complete syntax versus truncated, such as “the library is on the second floor” versus just “second floor.” Verbose greeting is realized by treating asking “How are you?” which is treated as an information-seeking question that requires a response and a reciprocal question, such as “I am doing well, thanks, you?” resulting in a multi-turn exchange. Positive and negative valences of verbosity across all dialogue turns are shown in table 7.
- *Hedging* corresponds to the number of tentative words (“maybe”, “perhaps”, etc.) or expressions of uncertainty (“I think”, “if I am not mistaken”), where positive valence of hedging corresponds to a larger number of tentative words or expressions of uncertainty (see table 8).
- *Alignment* corresponds to the choice of syntactic structures or words that mimic those of a conversation partner (*interlocutor*). Positive alignment corresponds to preference towards the choices of the interlocutor, specifically, higher overlap in syntactic structure and lexical choice. An example of a lexical alignment would be both interlocutors using the same synonym, such as “bathroom”, instead of one interlocutor using “bathroom” and the other “restroom.” An example of a syntactic alignment in the context of a greeting would be the pair of utterances “How are you doing?” and “I am doing well”, as opposed to a non-aligned pair “How are you doing?” and “Not bad.” The alignment stimuli across all dialogue turns are shown in table 9.
- *Formality* corresponds to a degree of social distance expressed in the language. Positive formality is realized via a choice of words and syntactic structures, specifically, terms of address that highlight social distance, demonstrate reverence, humble the speaker and elevate the addressee. In our case we use “sir” and “madam” as formal terms of address, polite markers such as “kindly (follow...)”, humbling devices “I am afraid” and “I have to apologize” (see table 10).

Since we are primarily interested in dialogue, the stimuli include fragments of dialogues consisting of 2 to 6 consecutive dialogue turns by two interlocutors. For each dimension of language variability and each of two valences (and a neutral control condition shown in Table 6) we present dialogue fragments corresponding to four dialogue acts of greeting, answer to a question, disagreement and apology. The

participant is asked to imagine that he or she is the visitor and the other interlocutor is described as a “female receptionist in her early 20s and of the same ethnic background” as that of the participant. The description of the occupation, age, gender and ethnicity of the interlocutor whose utterances the participant is asked to evaluate should provide a basic context and help avoid variability due to the implicit assumptions that subjects may make. An example of a web page presenting an item of the stimuli, including the description of the context, a dialogue fragment and the naturalness item, is shown in figure 24. Appendix A.1 contains an example page with personality scale questionnaire (figure 39) and the full set of stimuli (Tables 6–10).

Imagine that you are visiting a university building for the first time and see a female receptionist who appears to be in her early 20s and of the same ethnic background as yourself. The following conversation takes place between you and the female receptionist. (This text will not change from page to page, but the conversations below will be different.)

You	Receptionist
Good morning.	Good morning. How are you today?
I am doing well, thanks. You?	Very well, thank you. How's your family?
Everyone is doing fine, thanks. How about yours?	Mine is doing well too. How may I help you?

Do you agree that the Receptionist's utterance was natural?

Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>

Figure 24: A rendering of the web page that presents a positive verbosity of the greeting with the naturalness question. Symbol \bigcirc represents HTML radio button element.

4.2.4 Measures

We ran two versions of the study in parallel. In one version, participants were asked to evaluate the receptionist's utterances with respect to measures of personality using the ten-item personality question-

naire (TIPI, see [Gosling et al. \[2003\]](#)). In the other version, participants were asked to evaluate the receptionist's utterances with respect to their naturalness on a 7-point Likert scale. We used the naturalness item as a proxy for ethnicity, assuming that a behavior that is natural for a participant of ethnicity A, and not natural for a participant of ethnicity B, is a potential rich point. Although the naturalness item performed reasonably well (i.e. several stimuli were rated differently on naturalness by the participants of different ethnicities), in further studies we switched to more direct questions on ethnic attribution, hoping to measure perceived ethnicity in a more direct way.

The TIPI measure has been previously used in evaluating the personality of text fragments by [Mairesse and Walker \[2008\]](#). It is shorter than other personality inventories (notably, the 240-item inventory NEO-PI-R proposed in [Costa and McCrae \[1992\]](#)) and at the same time it retains a high degree of accuracy [[Gosling et al., 2003](#)], making it suitable for studies that require repeated measures of multiple stimuli.

In the naturalness version, we replaced the TIPI questionnaire with the question "Do you agree that the receptionist's utterances were natural?" A similar measure of naturalness has been previously used in evaluations of computer-generated speech by [Mairesse and Walker \[2010\]](#).

In summary, the naturalness version of the study had one question per stimulus, as opposed to ten questions per stimulus in the personality version. In an attempt to balance the workload of participants, we doubled the number of stimuli presented to each participant in the naturalness version, by presenting within one session the 24 conditions that correspond to two dimensions of linguistic variability.

4.2.5 *Subjects*

We recruited native speakers of American English from the United States and native speakers of Arabic from any of the countries of North Africa and Middle East with a significant fraction of Arabic-speaking population, namely Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestinian Territories, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen.

Since this is a pilot study, we did not have good estimates of the expected effect size. We judged that for an effect of 0.5, with power of 0.6 at significance level 0.05 we would need around 22 participants. We recruited participants via MTurk until each study condition (a combination of a dimension of linguistic variability, measure, and ethnicity of the participant) was successfully completed by at least 25 participants and at least 10 of each gender. Since tasks (or HITs, Human Intelligence Tasks, in Amazon's lingo) were requested in batches of 5 or 10, some conditions were completed by a few more subjects than

the minimum required. Due to the lower rate of participation, there were only 13-20 participants per each of the Arabic conditions. Nevertheless, the existing data has proved sufficient to draw a number of statistically significant inferences.

Upon completion of each HIT, participants received monetary reward as a credit to their MTurk account. Special measures were taken to prevent multiple participation of one person in the same study condition: the study website access would be refused for such a user based on the IP address, and MTurk logs were manually checked for repeated MTurk user names to detect logging into the same MTurk account from different IP addresses.

Overall, we noticed lower participation of workers from Arabic speaking countries as compared with workers from the United States. This led us to deploy a number of techniques, suggested by [Higgins et al. \[2010\]](#) to help promote the study among Arabic speaking MTurk worker community:

- We added descriptions in Arabic language and script to the titles of the HITs.
- We temporarily increased rewards for a subset of HITs to help promote the study among workers. Such an increase also pushes HITs towards the top of search results ordered by the amount of reward.
- We reposted HIT requests frequently (on a daily basis) to ensure high rating of the HITs in search results ordered by recency.

The demographics of the study participants by countries is shown in table 3. We counted each participant and study combination once, since a participant was allowed to take any subset of the 6 study conditions (4 conditions with the personality measure and 2 conditions with the naturalness measure). While we attempted to balance the conditions by gender, we observed a bias towards higher female participation among American workers and an opposite bias towards male participation among workers from the Arabic speaking countries. In total, there were 100 male and 55 female participants in the Arabic condition, and 63 male and 103 female participants in the American English condition. Interestingly, [Ross et al. \[2010\]](#) reported similar biases by sampling workers from the United States and India.

4.2.6 Results

We fitted linear mixed-effects (LME) models [[Pinheiro and Bates, 2000](#)] to the data and performed model selection using likelihood ratio tests. For the rationale on the model fitting technique, we refer the reader to Appendix A.2. The comparison of models fitted to explain the personality and naturalness scores (controlling for language and gender)

Language	Country	N
Arabic	Algeria	1
	Bahrain	1
	Egypt	56
	Jordan	32
	Morocco	45
	Palestinian Territory	1
	Qatar	1
	Saudi Arabia	5
	United Arab Emirates	13
	Total	155
American English	United States	166

Table 3: Distribution of study participants by country.

shows significant main effects of valence and dialogue acts for all pairs of personality traits (and naturalness) and linguistic features (see Tables 13–16). This means that averaging results over dialogue acts is not appropriate. Note that the effect of the participant's gender was not significant.

The results also show that for every personality trait (and naturalness) there is a linguistic feature that results in a significant three-way interaction between its valence, the native language, and the dialogue act. These results suggest that (a) for both language communities, every linguistic dimension is associated with every personality trait and naturalness, for at least some of the dialogue acts, (b) there are differences in the perception of every personality trait and naturalness between the two language communities.

To further explore the latter finding, we conducted a post-hoc analysis consisting of paired t-tests that were performed pairwise between the three values of valence for each combination of language, linguistic feature, and personality trait (and naturalness). Note, that comparing raw scores between the language conditions would be prone to find spurious differences due to potential culture-specific tendencies in scoring on the Likert scale: (a) perception of magnitudes and (b) appropriateness of the intensity of agreeing or disagreeing. Instead, we compare the language conditions with respect to (a) the relative order of the three valences and (b) the binarized score averages, namely whether the score is above 4 or below 4 (with scores that are not statistically significantly different from 4 excluded from comparison), where 4 is the neutral point of the 7-point Likert scale.

Selected results of the post-hoc analysis are shown in Figures 25–27. The most prominent cross-cultural differences were found in the

scoring of naturalness across the valences of the formality dimension (see Figure 25). Speakers of American English, unlike the speakers of Arabic, find formal utterances unnatural in greetings, question-answer and disagreement dialogue acts. Formal utterances tend to also be perceived as indicators of openness in disagreements and of conscientiousness in apologies by Arabic speakers, but not by American English speakers. Finally, hedging in apologies is perceived as an indicator of agreeableness by American English speakers, but not speakers of Arabic.

Interestingly, no qualitative differences across language conditions were found in the perception of extraversion and stability. This cross-cultural consistency can be interpreted in support of the previous findings that extraversion is one of most consistently identified dimensions (see, for example, Gill and Oberlander [2002]). It could also be possible that our stimuli were unable to pinpoint the extraversion-related rich points due to a choice of the linguistic dimensions, or particular wording. Larger variety of stimuli per condition, and an ethnography to identify potentially culture-specific linguistic devices of extraversion, could shed the light on this issue.

4.2.7 Discussion

Our analysis shows that the naturalness scale is a viable measure to identify cross-cultural differences in the perception of language. Personality scores too, while similar in many conditions, did show strong dependencies on the ethnicity of the grader. The biggest qualitative differences were:

- Apologies required manipulation of verbosity and formality to be scored positively on naturalness and on desirable personality traits of agreeableness and conscientiousness by Arabic speakers.
- High degree of formality differs dramatically on naturalness between American English and Arabic speakers.
- Lack of alignment has a dramatic effect on naturalness in both populations.
- Naturalness is not always correlated with personality traits.

The differences on personality scores may have multiple explanations. There is, for example, a possibility that the study stimuli reflect on the same “general interpersonal dimension of personality” [Scherer, 1972] for both groups of participants, but this general dimension is attributed differently to the Big Five personality traits, due to cross-cultural differences in implicit personality theories. Similar valid criticisms could be raised when comparing across cultures any self-reported measurements, including naturalness.

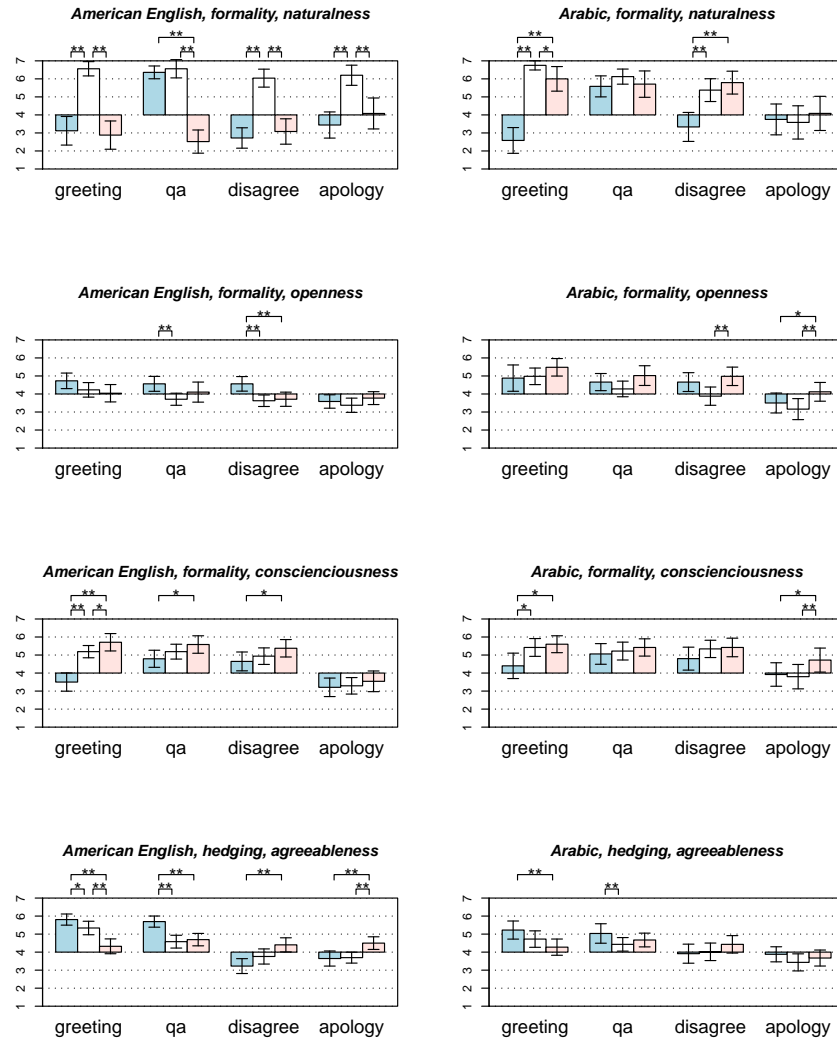


Figure 25: A subset of data comparing scores on the Big Five personality traits and naturalness as given by native speakers of American English (left half of the page) and Arabic (right half of the page). Blue, white, and pink bars correspond to negative, neutral, and positive valences of the linguistic features respectively. Dialogue acts listed along the horizontal axis are a greeting, question-answer pair, disagreement, and apology. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**) after Bonferroni correction.

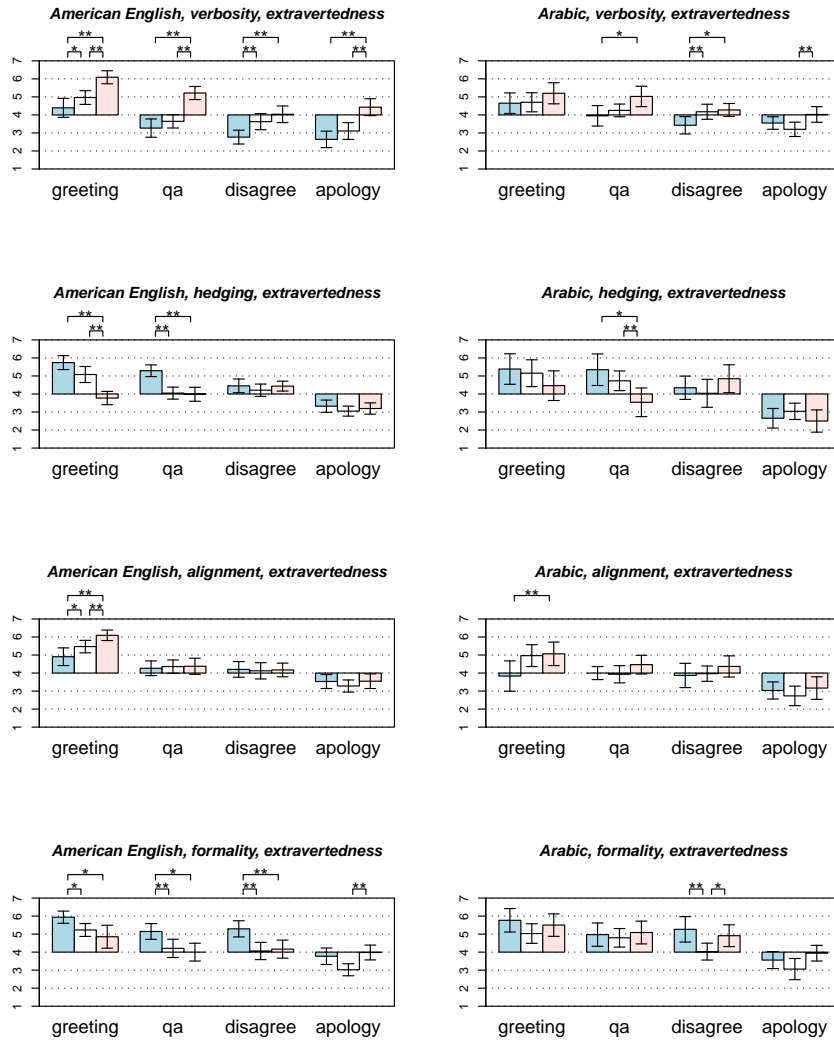


Figure 26: A subset of data (continued from Figure 25) comparing scores on the Big Five personality traits and naturalness as given by native speakers of American English (left half of the page) and Arabic (right half of the page). Blue, white, and pink bars correspond to negative, neutral, and positive valences of the linguistic features respectively. Dialogue acts listed along the horizontal axis are a greeting, question-answer pair, disagreement, and apology. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**) after Bonferroni correction.

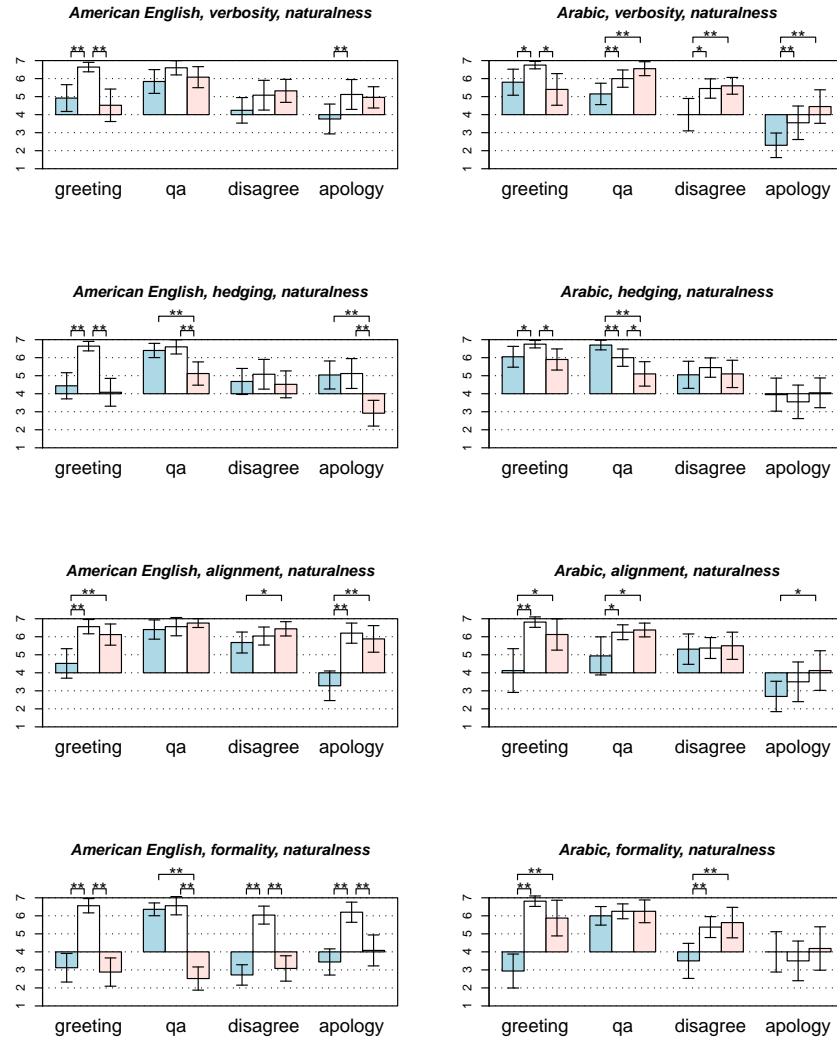


Figure 27: A subset of data comparing scores on naturalness as given by native speakers of American English (left half of the page) and Arabic (right half of the page). Blue, white, and pink bars correspond to negative, neutral, and positive valences of the linguistic features respectively. Dialogue acts listed along the horizontal axis are a greeting, question-answer pair, disagreement, and apology. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**) after Bonferroni correction.

Since the stimuli consisted of one dialogue per a combination of valence and linguistic dimension, observed effects may be due to idiosyncratic features of the particular dialogues. In that case, generalizations based on these results may not transfer well to different utterance realizations. The utterance realizations we used, while inspired by the related literature, were not sourced from the target populations. Naturalness scale has proven to lend to an easy interpretation, but its association with ethnic attribution is unclear. In Section 4.3.4 we will address these issues by evaluating ethnic attribution of concrete realizations (instead of attempting to generalize over linguistic dimensions) sourced from the corpus analysis presented in Section 3.2.

4.3 JOINT EVALUATION OF VERBAL AND NON-VERBAL BEHAVIORS

This section presents our main crowdsourcing study that aims at evaluating ethnic salience of a set of rich points and rich point candidates selected by performing corpus analysis and drawing upon intuitions from the related studies (Section 3.2). In the study, we attempt to address the deficiencies of the pilot study, presented in the previous section, by using an explicit ethnic attribution, sourcing the realizations of behaviors from our corpus analysis, and using stimuli that are similar to those used in the final study with the colocated robot (presented in Chapter 5).

The main goal of our crowdsourcing study is to jointly evaluate perception of verbal and non-verbal behaviors. We rendered the behaviors on the robot prototype, and used their videos as stimuli, to ensure the close relation between the stimuli of the final experiment and the online study. However, to be able to draw inferences about the influence of behaviors on perceptual (and, later, also objective) measures, we would like to be able to generalize over the robot's appearance. Another concern is that appearance can have an effect on ethnic attribution and on perceptual measures of on-screen agents (see, for example, [Baylor and Kim \[2004\]](#)). Similarly, voice features, such as prosody, may affect attributions of ethnicity (e.g. [Todd \[2002\]](#)) and personality (e.g. [Scherer \[1972\]](#); [Markel et al. \[1972\]](#)). Therefore we precede the main crowdsourcing study of this chapter by first selecting and evaluating the robot's appearance and voice.

4.3.1 *Selecting faces with high ethnic attribution*

To control for the ethnic attribution of the robot's face, we wanted to select four faces: two human-like, with strong ethnic attributions as AmE and Ar, and two robotic faces with low ethnic attribution. In this section, we describe the study on selecting the two human-like faces. All four faces, including the two robotic ones, are further evaluated Section 4.3.2. While it is not required, we were hoping to select faces that produce agreeable attributions by both of our subject populations: native speakers of American English residing in the US and native speakers of Arabic residing in one of the Arabic-speaking countries of North Africa and Middle East with Arabic as an official language.¹ Note, that since neither AmE nor Ar are uniform racial groups, instead of a clear dominant face we expect to see a distribution.

¹ We limit Ar participants to those with IP addresses from the following countries: Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestinian Territories, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen.

4.3.1.1 *Stimuli*

An artist² designed 18 human-like female face models (figure 28) that have identical 3d wireframe models but range along the following dimensions:

- 3 values of skin tone. Using the terminology from Fitzpatrick [1975], they are: dark (s₁), dark intermediate, or “olive” (s₂), and light (s₃).
- 2 values of eye color: brown (e₁) and blue (e₂), in terms of Martin-Schultz scale, described in Mackey et al. [2011].
- 3 values of hair color: black (h₁), brown (h₂) and blond (h₃), according to Fischer-Saller scale described in Hrdy [1978].

The participants were presented 2 pages of stimuli. In the first page, the participants were shown the complete set of 18 human-like faces in a random order and asked the following:

Put a checkmark under every face that looks like a person of your own ethnicity (in other words, a person who could be your close relative). You can mark multiple faces.

The second page had the same 18 faces in a generally different random order and, for AmE subjects, the task was as follows:

Put a checkmark under every face that looks like a person who is a native speaker of Arabic. You can mark multiple faces.

For Ar subjects, the task description was:

Put a checkmark under every face that looks like a person who is a native speaker of American English. You can mark multiple faces.

4.3.1.2 *Participants*

The participants were recruited using Amazon’s Mechanical Turk. The task description specified that we were looking for native speakers of American English residing in the US (AmE condition) or native speakers of Arabic from one of the set of the countries with Arabic as an official language (Ar condition). Participants were given a demographic questionnaire that asked (in English) their sex, age, race (based on the US census classification, AmE condition only), native

² Rich Colburn.



Figure 28: Human-like faces.

language and English proficiency (figure 47). We also asked participants to list up to three countries where they lived the most of their lives, with the number of years they lived in each.

In total, 53 speakers of American English (AmE) residing in the US, and 50 speakers of Arabic (Ar) residing in predominantly Arabic speaking countries participated in the study. Among AmE participants, 43 declared themselves as White, 7 as Asian, 2 as Black or African American. One participant did not provide an answer to the race question.

4.3.1.3 Results

On average, participants selected 4.9 images from each page ($SD = 3.2$). Figure 29 show votes for each face's attributions by both populations. Note the higher AmE attribution of light skin faces, compared with dark skin ones, likely due to the unbalanced participant pool in AmE condition. AmE and Ar subjects agreed on one image that resembled a native speaker of Arabic the most. There was less agreement on the appearance that resembles a native speaker of American English, as multiple images received high scores from AmE subjects. In the end, we chosen s2e1h1 (figure 28g, further referred to as face 1), favored by both AmE and Ar subjects as someone resembling a speaker of Arabic, s3e2h3 (figure 28r, further referred to as face 2), the face that was a clear favorite for Ar participants and that was among the highly rated by AmE participants as someone resembling a speaker of American English.

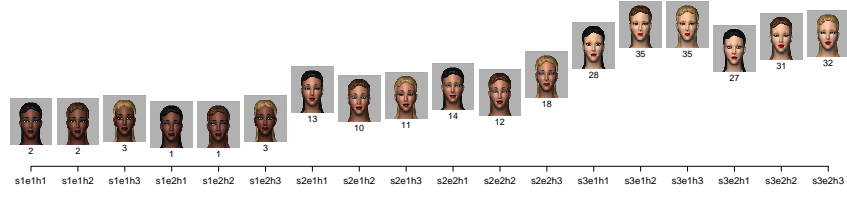
4.3.2 Scoring ethnic attribution of faces

4.3.2.1 Stimuli

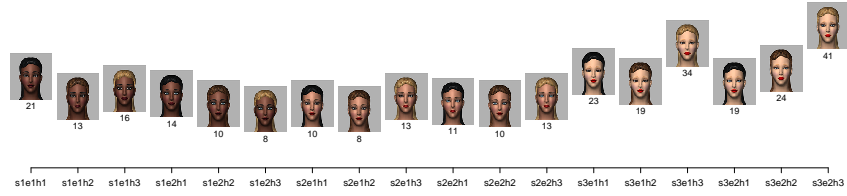
The same artist used the human-like face models to design two robotic faces: face 3 (figure 31c) and face 4 (figure 31d). We tested their ethnic attribution by giving the participants the following task description:

This image is used to represent an animated character of a receptionist. We ask you to guess that character's ethnicity, in terms of the character's native language. Please rate your guess.

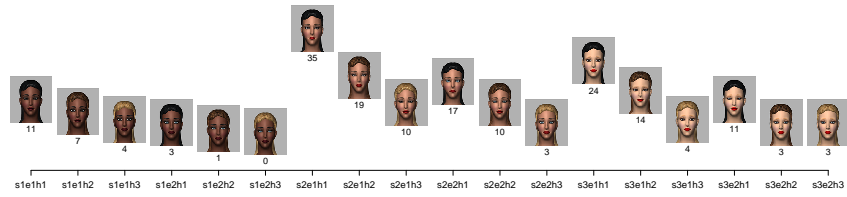
The participants rated their guesses on two 5-point semantic differential scale (Not American English—American English, Not Arabic—Arabic), that we deployed to be consistent with the Godspeed questionnaire that we used for the main crowdsourcing experiment (see Bartneck et al. [2009] for a detailed introduction of the Godspeed instrument). For all conditions, all textual descriptions (except for the demographic questionnaire) were also provided in Arabic (figure 30).



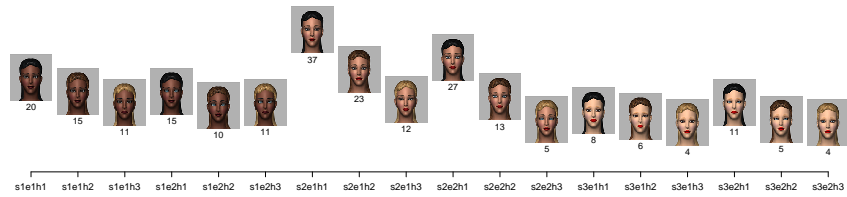
(a) Votes by AmE participants on face attribution to their own ethnicity.



(b) Votes by Ar participants on face attribution to AmE.



(c) Votes by Ar participants on face attribution to their own ethnicity.



(d) Votes by AmE participants on face attribution to Ar.

Figure 29: Distributions of votes on ethnic attribution of the 18 human-like faces. The horizontal axis labels denote skin tone, eye, and hair color.

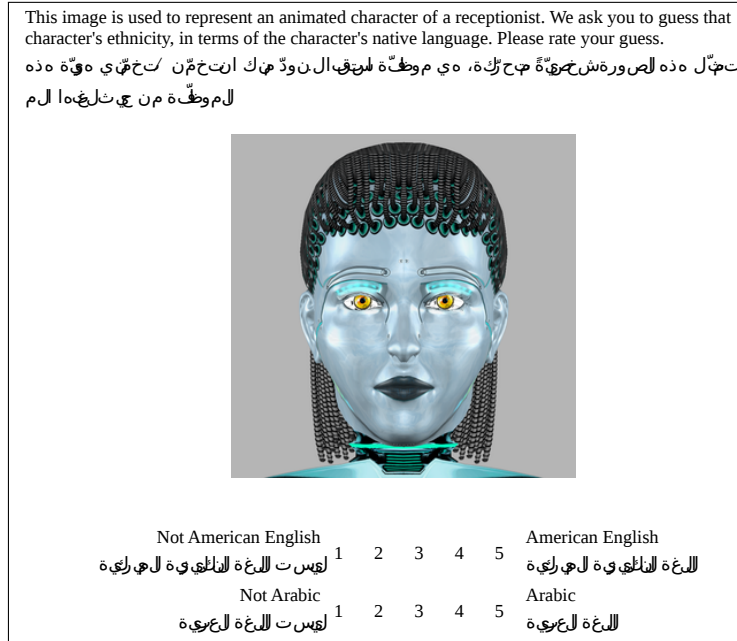


Figure 30: The stimulus for evaluation of the ethnic attribution of a face.

To verify that ethnic attribution of face 1 and face 2, selected from the set of 18 faces, would be possible without other faces readily available for comparison, we included those two faces in the random permutation of the four faces (one face per web page) in the experiment.

4.3.2.2 Participants

We recruited participants using MTurk, as described in Section 4.3.1.2. In total, there were 14 participants in the Ar condition and 17 participants in the AmE condition.

4.3.2.3 Results

Ar participants rated face 1 as more Ar, face 2 as more AmE. Face 3 and face 4 did not result in a significantly different ethnic attributions, with average attributions below the midpoint of the semantic differential scale.

AmE participants rated Face 2 as more AmE, but, somewhat unexpectedly, did not show significant differences in ethnic attribution of Face1. Face 3, however, was rated by AmE participants as more Ar, perhaps because of the braids.

4.3.2.4 Discussion

While the difference in attribution on face 1 in the scores of AmE participants is not significant, it is in the right direction. The unexpectedly close scores for this face, as opposed to the votes it received

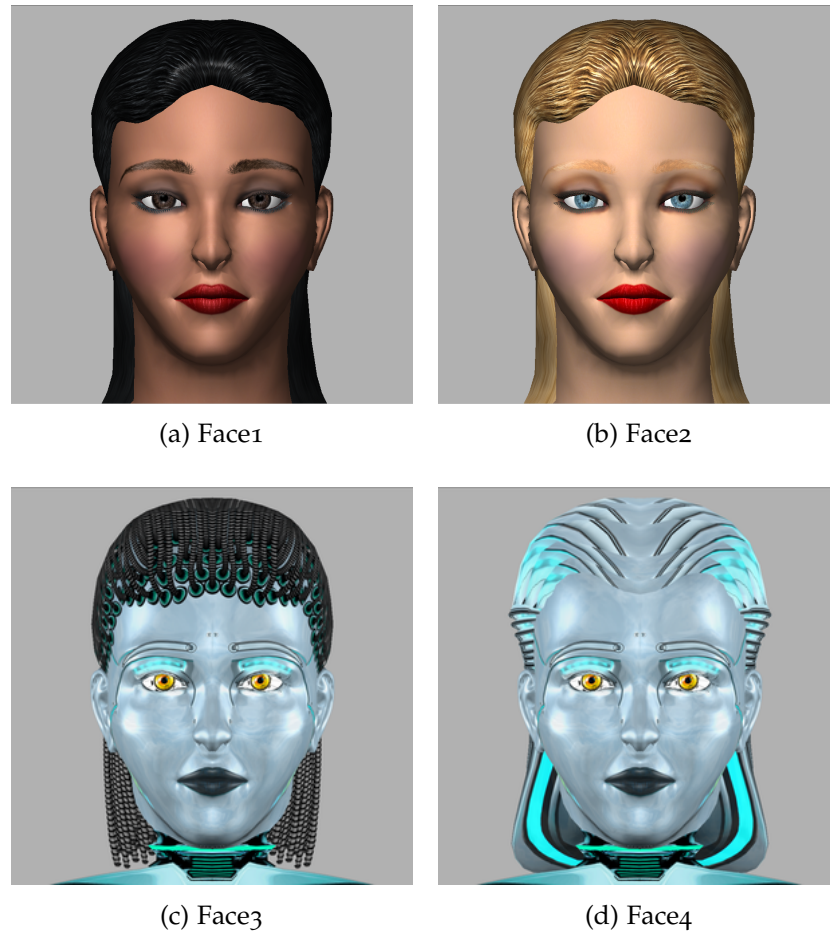
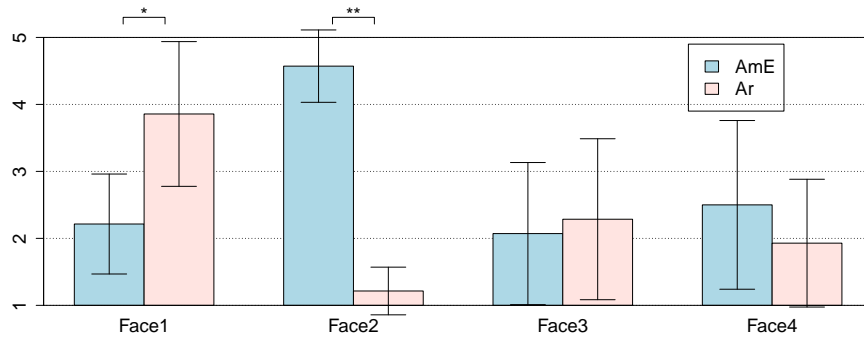
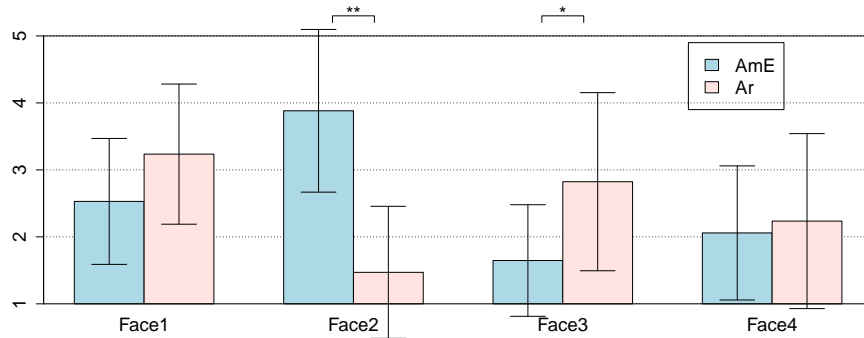


Figure 31: Human-like Face1 and Face2 were selected based on the large number of votes they collected as someone resembling native speakers of Arabic and American English, respectively. Face3 and Face4 are the robotic faces that elicit low ethnic attribution.



(a) Attribution of the four faces by native speakers of Arabic.



(b) Attribution of the four faces by native speakers of American English.

Figure 32: Ethnic attributions of the four faces. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**).

in the previous study of selecting faces from the set of 18, can be explained by the differences in stimuli and measure. In the earlier study, it was shown on the same page with 17 other faces, potentially biasing the AmE subjects towards stereotypical choices for the Ar appearance. The measure of selecting the face that may look like a close relative, on the other hand, biases away from the stereotypical choice for AmE appearance. As it happens, the stereotypical, in the view of AmE participants, Ar face 1, is scored, by AmE participants, as about equally likely to belong to a character who is a native speaker of Arabic or a character who is a native speaker of American English.

4.3.3 *Evaluating voices*

4.3.3.1 *Stimuli*

The robot prototype uses the Acapela text-to-speech software [[Acapela Group](#)]. We evaluated ethnic attribution of 6 English female voices: Heather, Laura, Lucy, Nelly, Rachel, and Tracy. Each web page of the test contained a widget that allowed the participant to replay the voice snippet reciting a count of integers from 1 to 10. The page had two questions scored on a 7-point Likert scale. The first question was:

Do you agree that it is likely that the person speaking in this audio has the same ethnicity as yours?

The second question depended on the reported native language of the participant. For AmE participants it read:

Do you agree that it is likely that the person speaking in this audio is a native speaker of Arabic?

For Ar participants, the second question read:

Do you agree that it is likely that the person speaking in this audio is a native speaker of American English?

4.3.3.2 *Participants*

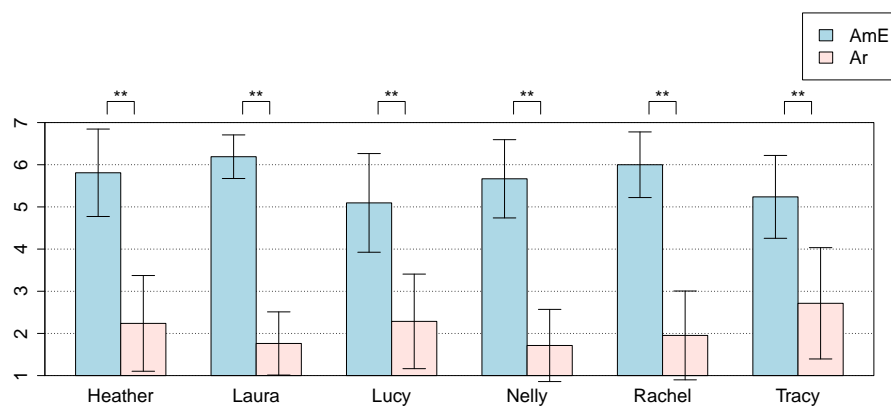
We recruited via MTurk 20 participants in AmE condition and 22 participants in Ar condition.

4.3.3.3 *Results*

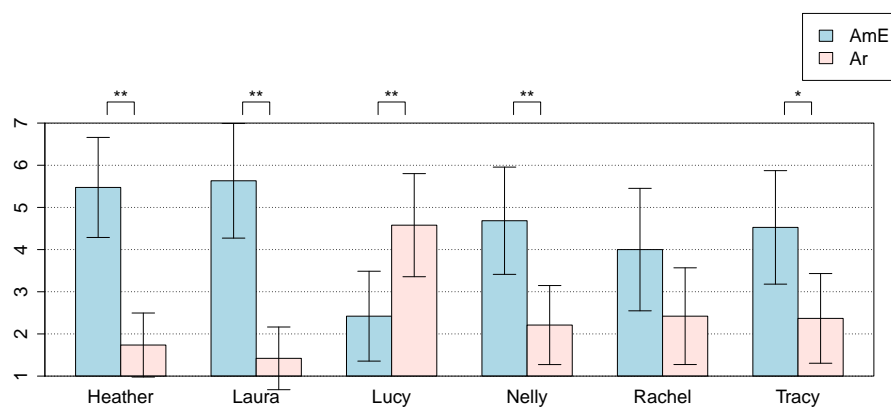
Ar participants attributed all six voices to AmE speakers, rather than Ar speakers. AmE participants attributes all voices to AmE speakers, except for the Lucy and Rachel voices. The British voice Lucy was attributed as more likely to belong to Ar, while the voice Rachel did not show a significant difference in ethnic attribution by AmE participants.

4.3.3.4 *Discussion*

Ideally, we would like to control for the ethnic attribution of voices as well, varying through a range of voices with various attributions. However, we were not able to obtain an English voice that would be attributed to a native speaker of Arabic by native speakers of Arabic. A larger study could generate custom voice for a text-to-speech software that satisfies the ethnic attribution requirements. Alternatively,



(a) Attribution of the voices by native speakers of Arabic.



(b) Attribution of the voices by native speakers of American English.

Figure 33: Ethnic attributions of the six Acapela voices. Error bars the 95% confidence intervals, brackets above the plots correspond to p-values of paired t-tests at significance levels of 0.05 (*) and 0.01 (**).

it is also possible to use canned recordings of human speech. For the remaining studies of this thesis, we selected the voice Laura, which has been used as the voice of Hala, has high attribution as AmE, and low attribution as Ar by both AmE and Ar participants. We discuss this, and other limitations of the studies, in Section 7.3.

4.3.4 *Evaluating videos of verbal and non-verbal behaviors*

Earlier, in Chapter 3, we argued for the need to evaluate rich points and rich point candidates on their ethnic salience. In this section, we describe our main crowdsourcing study that attempts to address this need.

The pilot study presented in Section 4.2 showed the feasibility of evaluating the perception of textual stimuli via online crowdsourcing. Appearance cues of ethnicity were selected and, together with the text-to-speech voices, were evaluated in a series of pre-studies, described in Sections 4.3.1 and 4.3.3. In this study, we combined appearance cues with the potential behavioral cues of ethnicity, that we identified from rich points and rich point candidates, to generate the videos of the stimuli. In the following section, we introduce the procedure for evaluating the perception of these stimuli.

4.3.4.1 *Procedure*

Ar and AmE participants recruited via MTurk were assigned to one of the 4 study conditions: (1) greetings, (2) gaze during direction-giving, (3) handling failure to provide answer, or (4) a combination of strategies for handling disagreement and politeness. The last study condition had two categories of stimuli combined to balance the load across conditions. After the participant completed the demographic questionnaire (figure 47), presented in English, he was taken through a randomized sequence of web pages each of which contained a description of a situation, an embedded video of the stimuli (hosted by YouTube), and the items of the Godspeed and ethnic attribution questionnaires (figures 40 and 41). All stimuli pages were presented in both English and Arabic.

Randomly within the sequence of stimuli we showed a validation page, designed to test whether the participant is able and is indeed viewing the video stimuli and can understand English (figure 42) or Arabic (figure 42), depending on the native language condition. All participants who completed the study were paid, but their responses to the validation questions were periodically checked. In this study, none of the participants failed the validation question.

4.3.4.2 *Stimuli*

The stimuli were presented to the participants recruited via MTurk as videos, each hosted by YouTube and embedded into a web page with the Godspeed questionnaire (see Bartneck et al. [2009]) for the description) and two questions on ethnic attribution (figures 40 and 41).

The robot's utterance were spoken in the voice Laura by the Acapela text to speech software, with the mouth of the robot's animated face

moving in synchrony with the spoken phonemes. Assumed user utterances were silent and shown as captions embedded into the video. The detailed description of the stimuli and the URLs for the videos are listed in Appendix B. Videos were pretested by 3 people to validate that the smiles and nods are recognized as such. All stimuli were shown with faces 1, 2, and 3. Face 4 was not used to reduce the workload on the study participants. Although, based on the results of Section 4.3.2, face 4 may be a better candidate for a face with a neutral ethnic attribution, this study was performed before the evaluation of ethnic attribution of faces (but after the study on selection of the faces, described in Section 4.3.1).

We studied 5 categories of stimuli: three of them were different realizations of dialogue acts of greetings, handling failure and handling disagreement, and two corresponding to gaze patterns and politeness markers during direction-giving. The stimuli behaviors were selected to match as closely as possible the original rich point (or rich point candidate) behaviors observed in our receptionist corpus (Section 3.2). Although we varied through all possible combinations of verbal and nonverbal behaviors for the greetings, that has proven to decrease the participant enrollment and the study completion rate, even though we tried to counterbalance the increased workload with a larger pay. Splitting such a large number of combinations from one stimuli category between multiple studies would result in the inability to perform complete within-subject comparison, possibly reducing the power of the analysis.

Because Reid said so.

Greetings. Utterances “yes, sir”/“yes, ma’am” (depending on the reported gender of the MTurk worker) and “hi” were presented in all possible combinations with physical robot head nod, the virtual head nod, broad smile, or no movements besides mouth movements related to speaking. The 8 combinations were presented with each of the three faces, for the total of 24 stimuli. The utterances were presented as parts of greeting exchanges:

Dialogue 1:

v: Hello
r: Yes, sir (Yes, ma’am)

Dialogue 2:

v: Hello
r: Hi

Direction giving. A fixed 4-step direction giving utterance was combined with 6 different gaze stimuli: (1) moving the gaze towards the destination at the beginning of the first turn and moving the gaze back towards the user at the end of the fourth turn, (2) gazing towards the destination at every turn for 1.2s, (3) gazing towards the destination at every other turn for 1.2s, (4) gazing towards the destination at every turn for 0.8s, (5) gazing towards the destination at every other

turn for 0.8s, (6) always looking towards the user (forward). In total, combined with the three faces, there were 18 stimuli in this study condition. Gaze transitions included both the turning of the physical screen, as well as virtual eye movement.

The verbal part of this stimuli is as follows:

v: Can you tell me where the library is?
 r: Library... Go through the door on your left.
 r: Turn right.
 r: Go across the atrium.
 r: Turn left at the hallway and you will see the library doors.

Handling failure to provide an answer. Two failure handling strategies were tested, both including a 0.6 second gaze to the left peaking at 0.2 rad angle. When presented across the 3 faces, this study condition contained 6 stimuli in total.

Dialogue 1, emphatic admission of failure, no explanation:

u: Can you tell me where the Dean's office is?
 r: I have no absolutely idea.

Dialogue 2, admission of failure, and an explanation:

u: Can you tell me where the Dean's office is?
 r: I don't know where it is, because I am new.

Handling a disagreement. Two dialogues that do not involve any nonverbal expression and vary only on how explicit the disagreement is. In an attempt to balance the workload of MTurk workers, these 6 stimuli (when presented across the 3 faces) were combined with the 6 stimuli of the politeness category into a single 12-stimuli study condition.

Dialogue 1, explicit disagreement:

u: Can you tell me where the cafeteria is?
 r: Cafeteria... Go through the door on your left,
 then turn right.
 u: Go through the door, then left?
 r: No, turn right.

Dialogue 2, implicit disagreement:

u: Can you tell me where the cafeteria is?
 r: Cafeteria... Go through the door on your left,
 then turn right.
 u: Go through the door, then left?
 r: Yes, turn right.

Politeness. The third direction-giving gaze condition (1.2s gazes towards destination ever other step) was varied with respect to the politeness of the robot's utterances. As we pointed out, these 6 stimuli (when varied across the 3 faces) were combined with the 6 stimuli for handling disagreement into a single 12-stimuli study condition.

Dialogue 1, no politeness markers (a direct style):

u: Can you tell me where the library is?
 r: Library... Go through the door on your left,
 turn right. Go across the atrium,
 turn left into the hallway and
 you will see the library doors.

Dialogue 2, with politeness markers (polite style):

u: Can you tell me where the library is?
 r: Library... Please go through the door on
 your left, then turn right. You may go
 across the atrium, then you may turn left
 into the hallway and you will see the
 library doors.

4.3.4.3 *Measures*

In this crowdsourcing study, instead of having a naturalness item, we attempted to measure ethnic attribution more directly, by giving the following prompt followed by the 5-point opposition scales Not American English—American English and Not Arabic—Arabic (figures 40 and 41).

The robot's utterances and movement copy the utterances and movements of a real person. Please rate the likely native language of that person:

We also used the Godspeed questionnaire (Bartneck et al. [2009]), as an additional perceptual measure, instead of the personality inventory, for the following reasons. First, we are going to use the Godspeed questionnaire as a proxy for perceptual measures of homophily in the study with a colocated robot, described in Chapter 5. Confirming that rich point candidates are indeed perceived differently across ethnic communities using the same perceptual measure can improve our chances of measuring consistent differences when the behaviors are perceived via a colocated robot. Second, unlike personality instruments, the Godspeed questionnaire is adapted and quite extensively tested on measuring the perception of robots. As we will see in the following sections that discuss the study results, we will have to resort to looking for perceptual differences on the Godspeed measures, as ethnic attribution measures prove not to reveal significant score differences for most of the behaviors.

4.3.4.4 *Results*

Each group of stimuli were scored as separate within-subject experiments to 16-35 workers in each of AmE and Ar communities. We tested the significance of the behaviors as predictors of the questionnaire scores by fitting mixed effects generalized linear models and

computing empirical confidence intervals. The results are discussed below. For the complete listing of significant results, we refer the reader to Appendix B.2.

Greetings. Table 17 shows significant perceptual measures for a linear mixed model that uses videos (each of which is a combination of face, verbal and nonverbal behaviors) as an independent variable. Table 18 shows significant perceptual measures for a linear mixed model that uses faces, and verbal and nonverbal behaviors as independent variables. Although AmE workers rated the verbal behavior “yes, sir”/“yes, ma’am” higher on likeability, intelligence, animacy and anthropomorphism, they gave it lower scores on AmE attribution. Surprisingly, the same verbal behavior scored higher on AmE attribution by Ar workers, but they did not consider it more likeable or intelligent. Smile and in-screen nod improved likeability for both AmE and Ar workers. However, AmE workers scored smile, physical nod, as well as face 3 as less safe, while Ar workers rated face 2 as less safe, and smile and in-screen nod as more safe nonverbal behaviors.

The comparison of the perceptual scores suggests that the greeting “Yes, sir”/“Yes, ma’am” elicits ethnic attribution, although differently for Ar and AmE workers. Physical nod is found less safe in various combinations by works in both groups. The data does not provide enough grounds to prefer smile over in-screen nod for any of the two ethnic groups.

Direction giving. Significant perceptual scores on direction giving stimuli are shown in Tables 19 and 20. AmE workers showed an effect of gaze 6 on increasing AmE attribution (primarily due to interactions with face 1 and face 3) and on decreasing Ar attribution. Both Ar and AmE workers scored all faces with gaze 6 lower on animacy and likeability. Face 2 had a main effect of lowering Ar attribution for Ar workers, and face 3 lowered AmE attribution for AmE workers.

In summary, Gaze 6 is positively associated AmE attribution and negatively associated with Ar attribution for native speakers of American English. There is no telling which of the gazes 1–5 may elicit stronger attribution of the robot character as Ar.

Handling failure to provide an answer. Significant perceptual scores on direction giving stimuli are shown in Tables 21 and 22. Ar workers rated the no-explanation strategy as less Ar and as more AmE. AmE workers found no-explanation strategy safer, except for face 2. Ar workers rated face 3 as less likeable.

In summary, no-explanation strategy appears to be a positive of AmE ethnicity and negative cue of Ar ethnicity to native speakers of Arabic.

Handling disagreement. Tables 23 and 24 show significant perceptual scores on disagreement strategy stimuli. The explicit disagreement behaviors were ranked as more intelligent by AmE workers and Ar workers, yet, less likeable by both. Ar workers, however, scored face 3 giving an explicit disagreement lower on likeability. No apparent cues of ethnicity were found.

Politeness. Polite markers during direction-giving were scored as more likeable by both AmE and Ar workers. Ar workers, however, also scored polite directions as more intelligent and animate. Again, no cues of ethnicity are apparent from the data.

Averaging over faces, behaviors, stimuli categories, and participant populations, robot characters were attributed as more likely to be AmE rather than Ar ($M_{\text{AmE}} = 3.90$, $M_{\text{Ar}} = 1.70$, $t(5638.87) = 28.07$, $p < 0.001$). This overall prevalence of AmE attribution is present within each of the stimuli categories and within both AmE and Ar participants. It should not be surprising, as all characters spoke English, with the same voice. Interestingly, few combinations of behaviors and faces showed differences in perception between the two participant populations. Also, there were few combinations of behaviors and faces that resulted in any shift in ethnic attribution of the robot characters.

4.3.4.5 Conclusions

Two main conclusions can be drawn. First, the ethnic attributions of the faces shown as images were generally not replicated by videos of the faces rendered on robots engaged in a dialogue. Second, since ethnic attributions rarely differ for these stimuli, other measures, such as Godspeed concepts, become more useful indicators of rich points. However, interpretation of the difference in Godspeed concept scores across subject groups is not that straightforward. For example, while AmE participants scored directions with polite markers higher only on likeability, Ar workers gave this combination higher scores also on intelligence and animacy. Does such differences in the number of positively associated perceptual measures make a behavior a good candidate for an ethnic cue? Even when ethnic attributions are significant, the subject groups may disagree on them. For example, the verbal behavior “yes, sir”/“yes, ma’am” has a negative effect on AmE attribution by AmE participants, but a positive effect on AmE attribution by Ar participants. This supports the idea that a behavior can be a cue of ethnicity for members of an out-group and not a cue (or,

in our case, a negative cues) when viewed by the members of the in-group.

Contrary to our hopes, not all categories of behaviors evaluated on their ethnic salience had definitive ethnic cues. Below, we summarize our arguments for choosing one or another behavior for ethnically salient robot characters described in Chapter 5. In cases when no behavior had significantly different ethnic attribution score, we relied on perceptual scores of Godspeed questionnaire. In some cases, we had to rely on our intuitions from the corpus data and the related work.

Greetings. AmE and Ar workers give the verbal behavior “Yes, sir”/ “Yes, ma’am” contradictory signs on its attribution as AmE. Nonverbal smile and in-screen nod both receive high perceptual scores on Godspeed measures, from both AmE and Ar workers. Since the source of “Yes, sir” and a nod is an expert Ar receptionist, we decided to use this combination in the characters designed to evoke attribution as Ar. The other combination, “Hi” and smile, will be used in characters designed to evoke AmE attribution.

Direction giving. Gaze 6 (no pointing gaze) is a positive cue of AmE ethnicity and a negative cue of Ar ethnicity, for AmE workers. This makes it a good choice to be used as a gaze behavior of a robot character designed to evoke AmE attribution. Both gaze 2 ($4 \times 1.2s$ pointing gazes) and gaze 4 ($4 \times 0.8s$ pointing gazes) were rated as more animate by Ar workers. We chose gaze 4 for the robot character designed to evoke Ar attribution because of a bias of Ar receptionists towards sub-second pointing gazes observed in our corpus analysis.

Handling failure to provide an answer. Since the strategy with no explanation is rated by Ar workers as a positive cue of AmE and a negative cue of Ar ethnicities, it makes it a good choice for a cue of AmE ethnicity. Another behavior, the explanation without lip stretch is going to be used as a cue of Ar ethnicity.

Handling disagreement. In one of the mixed models (Table 23), AmE perceived the explicit disagreement strategy as more intelligent. However, in the second model (Table 24), explicit disagreement had negative effect on intelligence for both AmE and Ar workers. We use the difference shown in former model, and the indirectness, ascribed to Arabic cultures in the related work, as justifications for selecting implicit disagreement as a cue of Ar ethnicity, and explicit disagreement as a cue of AmE ethnicity.

Politeness. Ar workers expressed positive perception of politeness markers on intelligence, likeability and animacy. AmE work-

ers considered polite strategy only as more likeable. AmE workers also considered it as less intelligent with face 3. Based on this comparison we select the polite strategy as a clue of Ar ethnicity, and the direct strategy as a cue of AmE ethnicity.

4.4 SUMMARY

This chapter presented our approach to evaluating the salience of rich points and rich point candidates as ethnic cues using online crowdsourcing. The first study of perception of textual stimuli gave us assurance that the target communities, in particular native speakers of Arabic, are available for recruitment via Amazon's Mechanical Turk. Using personality and naturalness measures, we were able to interpret scores for some of the dimensions of linguistic variability in terms of ethnic salience. Encouraged by these results, we conducted an online study that used as stimuli videos of both verbal and non-verbal behaviors. This time, we used perceptual measures that attempt to evaluate ethnic attribution via more direct questions, as well as the Godspeed questionnaire, tailored for human-robot interaction. The results of this study were harder to interpret. Few stimuli affected ethnic attribution scores, and even when they did, sometimes the attribution was different for different communities of participants. In other cases, stimuli affected subsets of perceptual measures of the Godspeed questionnaire, but often in the same direction for both groups of participants. Such cases were interpreted as follows. If sets of affected measures are comparable (one is a subset of another), the ethnicity of the participants for whom the behavior affected a larger set of perceptual measures was identified as the one for which the behavior is a potential ethnic cue. If sets of affected measures are identical or incomparable, we had to rely on the ethnicity of the behavior realization in our corpus data, and on our intuition from related studies.

In summary, while evaluating perception of the behaviors via crowdsourcing was not sufficient in itself for selection of ethnic cues, by drawing upon the corpus data analysis and the cultural models from related work, we were able to make an informed guess for all of our behavior categories. Whether these guesses were sufficient enough to be combined in robot characters that elicit ethnic attribution and homophily, remains to be seen. The study presented in the next chapter gives a partially positive answer to the ethnic attribution part of this question.

The analysis of the corpus of video dialogues and evaluation of behavior candidates via crowdsourcing allowed us to narrow down the set of rich points that are potential cues of ethnicity. We divided the most salient behaviors into two groups and presented them as two experimental conditions: behavioral cues of ethnicity 1 (BCE₁) and behavioral cues of ethnicity 2 (BCE₂). We hypothesize that these behavior conditions will affect perceived ethnicity of the robots (as we control for the appearance). We remind the reader that we continue to refer to the combination of a robot's appearance (varied by face) and behaviors (BCE) as a robot character.

What is a robot character, again?

Ethnic attribution family of hypotheses 1Ar and 1AmE: Behavioral cues of ethnicity will have a main effect on ethnic attribution of robot characters. In particular, BCE₁ will have a positive effect on the robot characters' attributions as Ar, and BCE₂ will positively affect the robot characters' attributions as AmE.

We also hypothesize that these behavior conditions will elicit homophily, measured as concepts of the Godspeed questionnaire (see [Bartneck et al. \[2008\]](#)), and objective measures of task performance (locating a destination on a map) and thanking the robot (see Section 5.4 for the detailed description of the measures).

Homophily family of hypotheses 2A–2G: A match between the behavioral cues of ethnicity expressed by a robot character and the participant's ethnicity will have a positive effect on the perceived animacy (2A), anthropomorphism (2B), likeability (2C), intelligence (2D), and safety (2E) of robot characters. Participants in the matching condition will also show improved ability to locate the destination on the map (2F) and they will thank the robot more (2G).

5.1 PARTICIPANTS

Adult native speakers of Arabic who are also fluent in English (Ar) and native speakers of American English (AmE) were recruited in Education City, Doha. Data corresponding to one Ar subject and 3 AmE subjects were excluded from the analysis due to the less than native level of language proficiency or deviations in protocol. After that, there were 17 subjects (7 males and 10 females) in the AmE condition and 13 subjects (7 males and 6 females) in the Ar condition. The majority of the Ar participants were university students (mean age 19.5, SD = 1.6), while the majority of AmE participants were university staff or faculty (mean age 26.4, SD = 9.2).



Figure 34: Experimental environment.

5.2 PROCEDURE

Experimental sessions involved one participant at a time. After participants completed a demographic questionnaire (Appendix C, figure 47) and the evaluation of their own emotional state (the safety part of Godspeed questionnaire, Appendix C, figure 48), they were introduced to the first robot character. The robot was located across the table from the participant, with the keyboard and the screen showing the typed text placed on the table (figure 34). This setup resembles the actual deployment setting of the Hala roboceptionist at the counter shared with a human guard and receptionist (Chapter 1, figure 1).

Participants were instructed that they would interact with four different robot characters by typing in English, and were asked to pay close attention to the destination directions as they would have to recall them later. Participants were informed that, although the robot would reply only in English, it acts as a receptionist character of a certain ethnic background: either a native speaker of American English

or a native speaker of Arabic speaking English as a foreign language. Participants were asked to pay close attention to the robot's behaviors and the content of speech as they would be asked to score the likely ethnic background of the character played by the robot. As we note in the discussion (Section 5.6), even such detailed instructions were sometime not sufficient for participants to accept the idea of a robot having an ethnicity or a native language.

Interaction with each robot character consisted of the three direction-seeking tasks, with the order of the four characters for each of the participants varied at random. Hence, each participant performed, in total, 12 direction-seeking tasks. Behavior conditions BCE1 and BCE2 varied at random within the following constraints: for each of the participants, pairs face 1–face 2 and face 3–face 4 should be assigned to different behavior conditions. This was done to balance the potential behavior effect between two machine-like characters (faces 3 and 4). Before each of the tasks, the participants were given verbal and written reminder of the protocol, namely, the dialogue acts that they are asked to type to the robot. The protocols for the three tasks are shown in Appendix C, figures 44–46. The robot's responses were chosen by an experimenter hidden from the user, to eliminate variability due to natural language processing issues. Post-study interviews indicated that participants did not suspect that the robot was not responding autonomously.

After the three direction-seeking tasks with a given character were completed, the participants were given a combination of the God-speed questionnaire and two 5-point opposition scale questions addressing the likely native language of the character acted by the robot (Appendix C, figures 49 and 50). All questionnaire items were presented in both Arabic and English. Participants were also asked to recall the steps of the directions to the professor's office in writing and by drawing the route to the professor's office on a map (Appendix C, figure 51).

5.3 STIMULI

The independent variables of the study were robot behavior (BCE1 or BCE2), robot's face (1–4), and destinations and corresponding routes for the first task (the professors' offices). The combinations of the conditions were mixed within and across-subjects.

Robot behaviors varied across dialogue acts of greeting, direction giving, handling disagreement and handling failure to provide directions. Combinations of verbal and nonverbal behaviors for each dialogue act and behavior condition are shown in Table 4. The robot's response to the user's final utterance (typically a variation of "bye" or

“thank you”) did not vary across conditions and was chosen between “bye” or “you are welcome,”¹ as appropriate.

Dialogue act	BCE ₁	BCE ₂
Greetings	“Yes sir (ma’am)” + virtual nod	“Hi” + open smile
Directions	0.8s pointing gaze at each step + politeness	Constant gaze on user
Disagreement	No explicit contradiction: “Yes, turn right”	An explicit contradiction: “No, turn right”
Handling failure	Admission of failure + brief gaze away + providing an excuse	Emphatic admission of failure + brief gaze away + lower lip stretch

Table 4: The manipulated verbal and nonverbal behaviors.

Each participant conducted 3 direction-seeking tasks with each of the four faces (12 tasks overall). Behavior conditions BCE₁ and BCE₂ were combined with the faces in such a way that face 1 and face 2 (as well as face 3 and face 4) were paired with different behaviors, as described in Section 5.2.

Task 1. In the first task, participants were asked to greet the robot, ask directions to a particular professor’s office, and end the conversation as they deemed appropriate. Four professor offices were used as destinations (one per each of the four robot characters that interacted with every subject). The directions to the professor’s offices started from the actual experiment site and corresponded to a map of an imaginary building. Each route consisted of 6 steps (e.g. “turn left into the hallway”) with varying combinations of turns and landmarks. The landmarks were used in the template “walk down the hallway until you see a [landmark], then turn [left or right],” once per route. Since we use the recall of directions as a measure of task performance, the directions were pretested to be feasible, but not trivial to memorize. The BCE₁ condition dialogue for location A (office of Prof. Adams) is shown below with example verbatim user utterances. Nonverbal behaviors are enclosed in parenthesis. Following the conversation analysis conventions, brackets indicate overlaps between verbal and nonverbal behaviors.

u: hello
r: [Yes, sir.]

¹ The response to thanks was, erroneously, “welcome” for the first 2 AmE and 6 Ar subjects. We controlled for this change in the analysis.

[(virtual nod)]
u: Prof Adams' office please
r: [Professor Adams office...]
[(0.8s gaze to the right)]
Please go through the door on your left
(0.1s pause)
[Turn left.]
[(0.8s gaze to the right)]
(0.1s pause)
[You may turn left into the hallway]
[(0.8s gaze to the right)]
(0.1s pause)
[Turn right at the end of the hallway]
[(0.8s gaze to the right)]
(0.1s pause)
[Walk down the hallway until you see]
[(0.8s gaze to the right)]
black chairs, then turn right.
(0.1s pause)
[Professor Adams office will be the]
[(0.8s gaze to the right)]
first door on your left.
u: thanks

The direction giving dialogue to the same destination in the condition BCE2 is as follows.

u: Hello.
r: [Hi.]
[(open smile)]
u: Can you please tell me how to get to professor Adams' office?
r: Professor Adams office...
Go through the door on your left
(0.1s pause)
Turn left.
(0.1s pause)
Turn left into the hallway
(0.1s pause)
Turn right at the end of the hallway
(0.1s pause)
Walk down the hallway until you see
black chairs, then turn right.
(0.1s pause)
Professor Adams office will be the
first door on your left.
u: Okay, thank you very much.

Task 2. In the second task, the participants had to ask the robot for directions to the cafeteria, and, after the robot gives directions, which are always “cafeteria... go through the door on your left, then turn right,” to imagine that they misunderstood the robot and ask to clarify if they should turn left after the door. The robot would then respond with the disagreement utterance, corresponding to conditions BCE₁ or BCE₂. An example dialogue in the BCE₁ condition is as follows.

u: Good afternoon.
 r: [Yes, ma'am.]
 [(virtual nod)]
 u: Can you please point me toward the cafeteria?
 r: [Cafeteria... Go through the door on your left,
 then turn right.]
 [(gaze to the right during the utterance)]
 u: Turn left?
 r: Yes, turn right.
 u: Okay, thank you.

The dialogue in the BCE₂ condition is shown below.

u: Hello.
 r: [Hi.]
 [(open smile)]
 u: How can I get to the Cafeteria please?
 r: [Cafeteria... Go through the door on your left,
 then turn right.]
 [(gaze to the right during the utterance)]
 u: Should I turn left after the door?
 r: No, turn right.
 u: Okay thanks!

Task 3. In the third task, the participants had to ask the robot for directions to the Dean’s office, for which the robot would execute one of the failure handling behaviors (see Table 4). The example dialogue in the BCE₁ condition is as follows.

u: Hi.
 r: [Yes, ma'am.]
 [(virtual nod)]
 u: Where can i find the Dean’s office?
 r: [I don’t know where it is ... because I am new]
 [(2.5s gaze to the left during the utterance)]
 u: okay thanks

The example dialogue in BCE₂ condition is shown below.

u: Hello.
 r: [Hi.]
 [(open smile)]
 u: Can you please tell me how to find the Dean's office?
 r: [I have absolutely no idea.]
 [(lower lip stretch and 2.5s gaze to the left during the utterance)]
 u: Okay, thank you then.

5.4 MEASURES

I didn't want to call it dumb.

— A Canadian participant on scoring the robot
 on Unintelligent—Intelligent scale, *A post-study interview*

Pre-study measures included a demographics questionnaire that asked for age, gender, race, native language, English proficiency, and three countries where the participant lived the longest (Appendix C, figure 47). Since people's definition of native language ability varies (see [Laitin, 2000] for an overview of the issue), the list of countries where the participant lived the longest gave us an objective measure that we used in addition to the participant's self-disclosure as a native speaker of American English or Arabic. In particular, we excluded data of 2 subjects whose did not live in the US or attended an American secondary school, but self-identified as native speakers of American English.

Pre-study questionnaires also included a self-assessed emotional state: the safety part of the Godspeed questionnaire. We used it to calibrate the participant's safety assessment of the robots during the experiment.

After participants performed 3 tasks with a robot character, they were given the full Godspeed questionnaire, assessing perceived animacy, anthropomorphism, likeability, intelligence, and safety of the robot character (Appendix C, figure 49). They were also asked to assess the likely ethnicity of the character that the robot plays, using two items with the same 5-point opposition scale as the Godspeed questionnaire: Not American English—American English, and Not Arabic—Arabic (Appendix Sec:App3Perculture6, figure 50). The questionnaire was presented both in English and Arabic. While the translation of the questionnaire into Arabic that we used ² was done without *back translation* (namely, the validation step of translating back from Arabic into English), as advised by Bartneck et al. [2009], internal consistency of the items within both pools of subjects (all Cronbach's alpha values are greater than 0.7) gives us some degree of confidence in the adequacy of the translated instrument.

² The translation is contributed by Micheline Ziadee.

We have asked the participants to recall the directions given to them in task 1 in writing (Appendix Sec:App3Perculture6, figure 50) and by marking the route on the map (Appendix Sec:App3Perculture6, figure 51). We pre-tested the directions and the map before the study to ensure that recall of the directions and drawing the route to the destination on the map were feasible, but not too easy. The goal of the direction recall and map task are two-fold. First, they serve as a distractor from the actual purpose of the study, thus, potentially reducing the subject bias towards the study goals. Second, they provide objective measures of task performance. In the present analysis, we used the success or failure of locating the destination on the map as a binary measure of the task performance.

Another objective measure that we use is the number of times participants thanked the robot characters. Previous work suggests that thanking, as well as greetings and farewells, may be related to the user's social attitude [Makatchev et al., 2009; Makatchev and Simmons, 2009; Lee et al., 2010b]. We contend that a hypothetical association between ethnic congruence and social attitude could be considered as a facet of homophily. In the present study, thanking the robot was not required of the users, while greeting and ending the conversation was part of the instructions (Appendix Sec:App3Perculture6, figures 44–46). Consequently, we use the number of times a participant chose to thank each robot character as an objective measure of social attitude. Multiple thanking instances within the same dialogue were ignored, resulting in each of the four robot characters encountered by a participant being thanked 0–3 times.

Finally, after the participant interacted with each of their 4 assigned robot characters, they were asked to participate in a short post-study interview. In the interview, the experimenter asked participants for their general impressions and whether they thought the robot was responding autonomously. While none of the participants expressed doubts in the autonomy of the robot, some participants expressed difficulties in interpretation of the ethnic attribution questions. We discuss this issue in Section 5.6.

5.5 RESULTS

Godspeed scales showed internal consistency with Cronbach's alpha scores above 0.7 for both AmE and Ar participants. All the p-values are given before correction for 2-hypothesis familywise error for hypotheses 1Ar and 1AmE, or 7-hypotheses familywise error for hypotheses 2A–2G. For detailed results we refer the reader to Appendix C.2.

5.5.1 *Main effect of robot behaviors on ethnic attribution. Hypotheses 1AmE and 1Ar*

5.5.1.1 *Attribution hypothesis 1Ar.*

The data does not indicate a main effect of BCE or faces on Ar attribution. However, a stepwise backward model selection by Akaike Information Criterion (AIC) suggests a negative effect of the interaction between face 4 and BCE2 on attribution ($F[3, 100] = 3.44$, $p = 0.020$). The significance of the interaction terms between faces and BCE is confirmed by comparing linear mixed effects models with subjects as random effects using the likelihood ratio test: $\chi^2(3, n = 120) = 9.48$, $p = 0.024$. The 95% highest posterior density (HPD) intervals for the coefficient for the interaction between BCE2 and face 2 is $[-2.14, -0.01]$ and for the coefficient for the interaction between BCE2 and face 4 is $[-2.07, -0.04]$.

While there is no evidence of a main effect of the faces, the results suggest that with respect to the attribution of the robot characters as native speakers of Arabic the participants were sensitive to manipulation of BCE only for the robot characters with faces 2 and 4. The situation is evident from the comparison of the means shown in figure 35. Instead of performing pairwise post-hoc tests that would not take into account the random effects of the subjects, we fitted mixed effects models for particular faces. Focusing on the conditions with face 2 and 4 yields the effect of the BCE term with $\chi^2(1, n = 59) = 7.89$, $p = 0.005$ and the entirely negative 95% HPD interval for BCE2 $[-1.19, -0.15]$. Focusing on BCE2 conditions, on the other hand, yields the effect of faces with $\chi^2(3, n = 60) = 9.60$, $p = 0.022$, but all 95% HPD intervals for the faces trap 0.

The analysis does not support the hypothesis of that varying behavioral cues of ethnicity between BCE1 and BCE2 affects the robot characters' attribution as Ar. This is because BCE2 increased attribution as Ar for characters with faces 1 and 3, and decreased attribution as Ar for characters with faces 2 and 4. There is no evidence of the main effect of the faces.

5.5.1.2 *Attribution hypothesis 1AmE.*

The likelihood ratio test for the mixed effects models supports the main effect of BCE on the robot characters' attributions as a native speaker of American English, $\chi^2(1, n = 120) = 5.44$, $p = 0.020$. Mean scores of the robot's attribution as AmE are $M_{BCE1} = 3.62$ ($SD = 1.21$) and $M_{BCE2} = 4.00$ ($SD = 0.92$). Figure 36a shows mean attribution scores for each of the robot characters. The significance of the interaction between BCE and gender is $\chi^2(1, n = 120) = 4.14$, $p = 0.042$. Controlling for an additive face term, the significance of BCE is $\chi^2(1, n = 120) = 4.88$, $p = 0.027$ and the significance of BCE

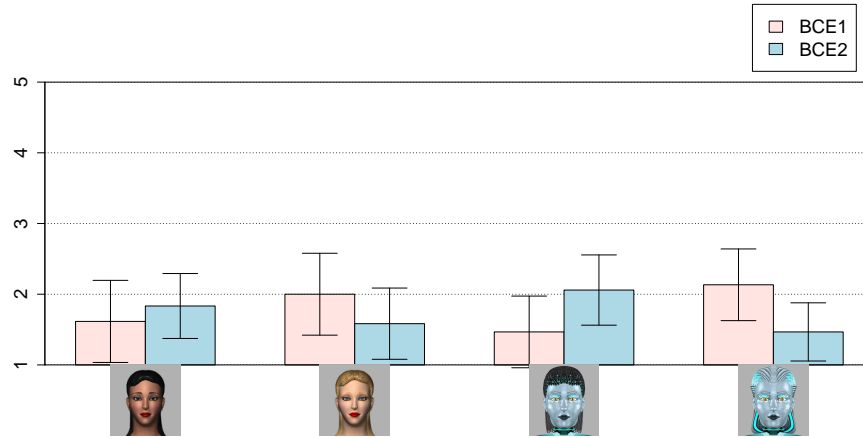


Figure 35: Score means on attribution of the robot characters as a native speaker of Arabic. Brackets correspond to 95% confidence intervals. This plot is for visualization only, as direct pairwise comparison would not account for subject effects.

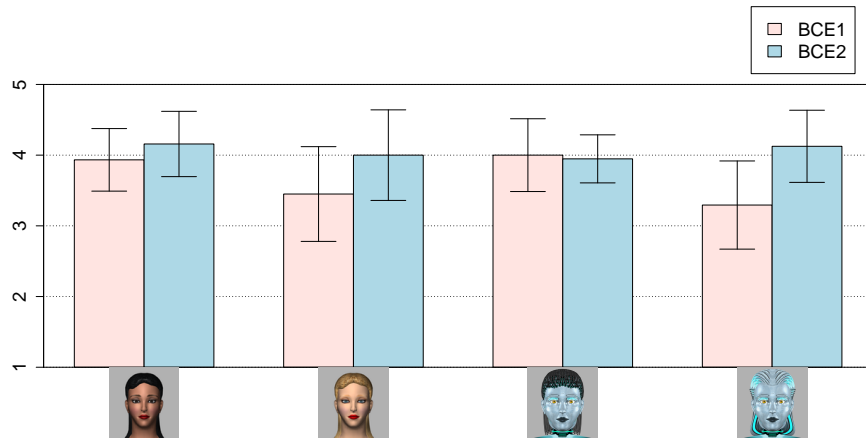
and gender interaction is $\chi^2(1, n = 120) = 5.02, p = 0.025$. The HPD intervals for the model's coefficients that do not trap 0 are $[0.20, 1.15]$ for BCE2 and $[-1.41, -0.02]$ for the interaction term between BCE2 and males.

The negative effect of interaction between males and BCE2 suggests that the behavioral cues of ethnicity would have a stronger effect among female subjects. Indeed, testing for the significance of BCE for female subjects yields $\chi^2(1, n = 64) = 8.74, p = 0.003$. The effect of BCE on male subjects does not suggest significance. This situation is evident from figures 36b and 36c. We note again, that post-hoc tests that would not account for a random effect of the subject are not appropriate here, because no subject had two behavior conditions with the same face of the robot character.

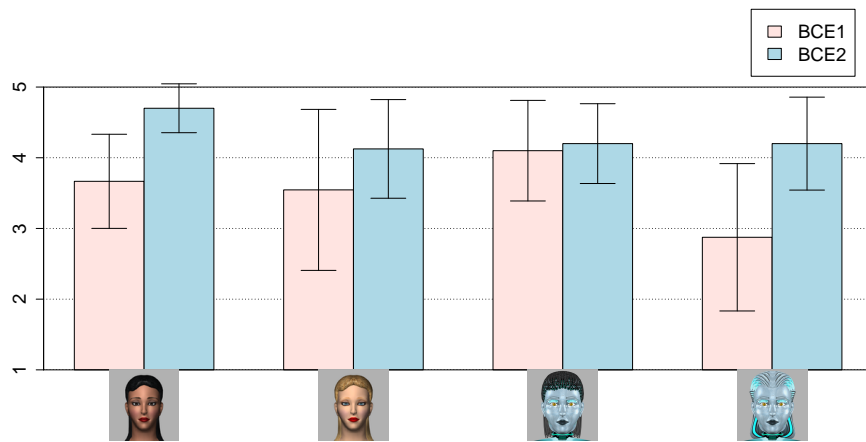
The analysis supports the hypothesis that varying behavioral cues of ethnicity from BCE1 to BCE2 has a positive effect on the attribution of the robot characters as native speakers of American English. This effect is more evident among females. There is no significant evidence of a main effect of the faces.

5.5.2 Homophily effects. Hypotheses 2A-G

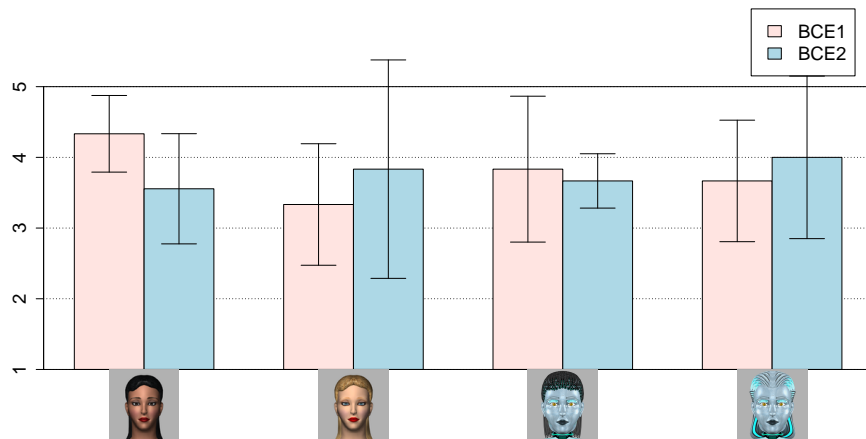
Tests of the interaction effects between the behavioral cues of ethnicity and the participant's native language do not support any of the homophily hypotheses. Further exploratory analysis showed the following associations. For the sake of readability, we will omit non-essential values of the statistics, referring the adventurous reader to Appendix C.2 for further details.



(a) Female and male participants.



(b) Female participants only.



(c) Male participants only.

Figure 36: Score means on attribution of the robot characters as a native speaker of American English. Brackets correspond to 95% confidence intervals. These plots are for visualization only, as direct pairwise comparison would not account for subject effects.

5.5.2.1 *Animacy*

BCE2 characters were rated as more animate $\chi^2(1, n = 120) = 6.00$, $p = 0.014$. Closing the interaction with “You are welcome” (the unintended independent variable) decreased animacy score $p < 0.0008$.

Within ethnic groups, AmE participants exhibited negative associations with animacy for face 4, $p = 0.018$, the closing “You are welcome,” $p = 0.019$, and positive associations with animacy for the combination between BCE2 and face 4, $p = 0.052$. The negative animacy of the robot with face 4 is mostly due to AmE female scores, as AmE males considered faces 2 and 4 as more animate ($p = 0.081$ and $p = 0.089$, respectively).

5.5.2.2 *Anthropomorphism*

AmE on average scored the robots as less anthropomorphic $p = 0.025$. The closing “You are welcome” also decreased perceived anthropomorphism, $p = 0.093$.

Within ethnic groups, AmE males gave lower scores on anthropomorphism on average, $p = 0.005$, in particular with face 1. Face 4 with BCE1 was scored lower on anthropomorphism, $p = 0.070$. Interestingly, AmE scored location D condition low on anthropomorphism, $p = 0.054$. This should remind the reader to treat these exploratory results as such.

No significant associations with anthropomorphism were found with model selection for Ar participants.

5.5.2.3 *Likeability*

Main effects associated with likeability are the closing “You are welcome” (negative, $p < 0.0004$), and the combination of BCE2 and face 4 (positive, $p = 0.025$). The positive effect of BCE2 and face 2 is present within AmE population ($p = 0.0547$) and is not significant within Ar participants. The negative effect of the closing phrase “You are welcome” is present within both participant groups.

5.5.2.4 *Intelligence*

The closing “You are welcome” is considered less intelligent ($p = 0.008$). This is mostly due to Ar subjects, with the effect present when fitting within the Ar group only $p = 0.025$.

Within ethnic groups, AmE participants rated the combination of face 3 and BCE2 as less intelligent, $p = 0.075$. AmE males rated face 3 as less intelligent, $p = 0.064$. Ar did not show any associations with the intelligence score, except for the reported negative association of the closing “You are welcome.”

5.5.2.5 *Safety*

AmE generally gave lower scores on safety, $p = 0.036$. Males also scored the robots lower on safety, $p < 0.0001$. The main effect of the gender is mostly due to Ar males $p < 0.0001$. The gender effect is confirmed by confidence intervals of the linear mixed-effect model with subjects as a random effect. The 95% HPD interval for the coefficient of the male term is $[-2.707, -0.715]$.

Fitting the models within the groups confirms that AmE participants gave higher safety scores to the combination of BCE2 and face 4, $p = 0.026$. Ar males, compared with Ar females, scored all the robots considerably less safe, $p < 0.0001$. These within-group trends are confirmed by fitting the safety scores adjusted by the pre-study questionnaire.

5.5.2.6 *Locating the destination on the map*

Task performance, in terms of success or failure of locating the professor's office on the map, did not show any significant associations with the independent variables. Fitting a linear model with the additive terms for destination and the closing shows a trend towards more success in locating the office C (Prof. Coopers). While there is a chance that the destination C genuinely had easier directions, it is more likely an artefact of the study design (lack of balance) and analysis. For example, although the order in which destinations were presented was randomized, the sample average was biased towards destination A shown the earliest (mean order $M = 1.07$), followed by destination B ($M = 1.56$), destination C ($M = 1.63$), and destination D ($M = 1.73$). Although it is possible that the participants had some training effect resulting in better performance for later locations, the data does not support this explanation. Specifically, fitting a linear model with the order of the robot character as an explanatory variable does not find a significant association with the task performance.

5.5.2.7 *Thanking*

Ar males thanked the robots more than Ar females ($p < 0.0001$), while AmE males thanked the robot less than AmE females did ($p < 0.0001$), except for face 2 ($p = 0.007$).

There was no effects of faces on behaviors on thanking. This can be interpreted as the failure of the appearance and behaviors to affect the social disposition of the participants.

5.6 DISCUSSION

It can be proven that most claimed research findings are false.

— John P. A. Ioannidis,

Why Most Published Research Findings Are False [Ioannidis, 2005]

The effect of behaviors on the perceived ethnicity of the robot characters was more evident in the attributions as a native speaker of American English rather than as a native speaker of Arabic. The fact that the behaviors did have an effect on attribution of the robot characters as Ar with faces 2 and 4 suggests that the manipulated behaviors were just not strong enough to distinguish between native and non-native speakers of American English, but also had relevance to the robot's attribution as Ar. The collocation of the questionnaire items for AmE and Ar attributions, however, could have suggested to the participants that these two possibilities are mutually exclusive and exhaustive. The correlation between the attribution scores is moderate, albeit significant, $r = -0.44$, $p < 0.0001$. An improved study design could evaluate only one of the attributions per condition or per participant.

The effect of behaviors on the attribution of the robot characters as AmE was evident only among female participants. A possible gender difference in perception of appearance, verbal and nonverbal cues may be a cause (for example, see [Hall et al., 2000] for an overview of the gender differences in expression and decoding of nonverbal cues of emotions). An interaction between the perceived gender of the robot character and the participant's gender could be at play as well. A future study could address this issue by varying the gender of the robot characters.

While behaviors may be stronger cues of ethnicity than faces for robot characters, there is evidence supporting a potential of complex interactions between the two. One explanation for the lack of the main effect of faces is a potentially insufficient degree of human likeness of our robot prototype. Post-study interviews indicated that many participants had difficulties in interpreting the ethnic attribution questions in the context of the robot, given that it spoke only in English. For example, one participant said "I am not sure which one was Arabic, because it always spoke in English." It would be interesting to replicate the study with more anthropomorphic robots that could, potentially, more readily allow attributions of ethnicity, such as Geminoid [Ishiguro, 2005].

Other limitations of this study include the fixed voice, with a strong attribution as AmE by participants of both ethnic groups, and the typed entry of user utterances. Some of the participants missed the robot's nonverbal cues as they were focusing their gaze on the keyboard even after they finished typing. We selected typed, rather than spoken, user input out of the concern that participants would not buy into the idea of the robot's autonomy, especially since some of them were familiar with the prototype (although with a different face) as a campus robot receptionist that relies on typed user input. This con-

cern may be less relevant for uninitiated participants, in which case an improved Wizard-of-Oz setup could rely on spoken user input.

Only AmE participants' data, only for some of the faces, was consistent with a possibility of behavior-associated ethnic homophily. The possible reasons for the almost complete lack of any homophily effects in our data include (a) the highly international environment, with most of the participants being students or faculty of Doha's American universities, and (b) a potentially low sensitivity of God-speed items and the chosen objective measure to the hypothesized homophily effects. It is also possible that the effect on ethnic attribution, while significant, was not large enough to trigger ethnic homophily. There is, however, evidence that in human encounters ethnically congruent behaviors can trigger homophily even when the ethnicities do not match (e.g. [Dew and Ward, 1993] and literature on cultural competence).

The study also provides a support for the consistency of the God-speed questionnaire, as well as its first known translation into the Arabic language.

Finally, as the epigraph to this section suggests, all research findings that are based on statistical analysis of data should be taken with a grain of salt. The reasons why the findings of an experiment may be wrong include hidden and apparent biases [Ryan et al., 2012; Wasserman, 2012] and an over-reliance on significance testing [Ioannidis, 2005]. Since every study has its biases, the best way to reduce them could be by replicating the study across different scenarios and participant pools. Such replications comprise one of the directions of the future work. We addressed the second possible culprit, significance testing, by conducting the exploratory analysis using both model selection and confidence intervals. Following Wasserman [2012], we remind the reader that "published findings are considered 'suggestions of things to look into,' not 'definitive final results.'"

5.7 CONCLUSIONS

Achieving ethnic homophily between humans and robots has the lure of improving a robot's perception and user's task performance. This, however, had not previously been tested, in part due to the difficulties of endowing a robot with ethnicity. We tackled this task by attempting to avoid overly obvious and potentially offensive labels of ethnicity and culture such as clothing, accent, or ethnic appearance (although we control for the latter), and instead by aiming at evoking ethnicity via verbal and nonverbal behaviors. We have also emphasized the robot as a performer, acting as a character of a receptionist.

Our experiment with a robot of a relatively low human likeness shows that we can evoke associations between the robot's behaviors and its attributed ethnicity. Specifically, we found an effect of behav-

iors on the attribution of the robot characters as a native speaker of American English by female participants, while the effect of behaviors on the attribution of the robot characters as a native speaker of Arabic was only evident for faces 2 and 4. Although we did not find a convincing evidence of ethnic homophily, we believe that the suggested pathway can be used to create robot characters with a higher degree of perceived similarity, and better chances of evoking homophily effect.

This study is a culmination of a sequence of studies that selected and evaluated behaviors that are potentially useful for the design of an ethnically salient robot character. We posited that the character design that aims at an ethnic attribution through its behaviors would benefit from a wide range of ethnically salient behaviors applicable in a multitude of dialogue acts. Mining for such behaviors requires a large set of behavior candidates. This problem has been approached before in the design of on-screen agents by mining cross-cultural corpora for maximally distinctive behaviors. The maximally distinctive behaviors, however, may not necessarily be the most ethnically salient ones. This difference necessitates an additional step that evaluates ethnic salience of the behavior candidates. For this sequence of studies, we performed this step using crowdsourcing with the participants recruited based on the native language and the country of residence.

The methodology of selecting candidate behaviors from qualitative studies and corpora analyses, evaluating their salience via crowdsourcing, and finally implementing the most salient behaviors on a physical robot prototype is not limited to ethnicity, but can potentially be extended to endowing robots with other aspects of human sociodemographic identities. In such cases, crowdsourcing does not only help to alleviate the hardware bottleneck inherent in HRI but can also facilitate recruitment of study participants of the target sociodemographic identity.

It doesn't crash anymore :-) but also doesn't work :-(

— Reid Simmons, *A note on a keyboard*

In this chapter, we describe the software that has been used to produce the stimuli for the studies described in Chapters 4 and 5 by implementing the behavioral and appearance cues of ethnicity on a robot. The hardware of the robot prototype has already been introduced in Section 1.1.1.

6.1 THE SOFTWARE ARCHITECTURE

The software architecture of Hala 2, the next version of Hala, consists of a number of modules that communicate via IPC [Reid Simmons, 2012] or sockets. The modules can be grouped into those that process sensor data and user input, the interaction manager (IM), the information retrieval modules, and the expression system. We briefly review the modules below and provide a more detailed overview of the Interaction Manager, which is the author's main contribution to Hala 2, in Section 6.2.

6.1.1 Sensor and user input processing modules

Hala 2 is equipped with a SICK laser scanner that generates a planar depth map of the robot surroundings. This map is used by a particle filter tracker to detect and track individual people. The proximity of each tracked person is passed to the IM, allowing the robot to engage passers by.

The typed user input to Hala 2 is passed to syntactic and semantic parsers and then to the IM. The IM, in turn, has an option to additionally process the user's utterance via the rule-based dialogue manager AINE [Ainebot]. The syntactic parser is a statistical parser that produces Stanford dependency graphs [de Marneffe et al., 2006]. The semantic parser is a trainable parser that attempts to find a best-matching example template, in a nearest-neighbour fashion.

6.1.2 Information retrieval modules

Since Hala 2 acts as a receptionist, it has to be able to answer information seeking questions, in particular, weather and campus directions. These information retrieval modules are invoked upon a request from the interaction manager and perform relevant queries of online and database resources. After the data is obtained, IM passes it on to natural language generator modules that insert it into utterance templates, returning the final surface realization of the robot's utterance. For the purposes of this thesis, the utterances were hard coded in IM recipes.

6.1.3 Expression system

6.1.4 Facial expression generation

The facial expressions are generated by driving a Blender model between key frames, using CONCEPT open source software [Delaunay and Greeff, 2012]. The animations are specified in Python, as lists of key frames. Each key frame is a set of triplets:

```
('<Action Unit code>', <magnitude>, <attack time>)
```

Table 5 provides the codes (based on FACS, introduced by Ekman and Friesen [1978]) for the action units that are used to generate non-verbal expressions for this study.

For each key frame, *magnitude* is the degree of activation of the specific action unit, generally on a scale from 0 to 1. *Attack time* is desired duration of the linearly interpolated animation between the current values of action units and the key frame. The key frame is considered attained when all action units of the key frame reach their specified positions. Once this happens, the system moves on to execute the next key frame in the list. For example, the animation

```
(
  (
    ('05', 0.45 0.1), ('06', 1.0, 0.6)
  )
  ('05', 0.3, 0.3)
)
```

specifies the first key frame with the upper eyelids at magnitude 0.45 to be attained within 0.1 s and with the cheek raiser at magnitude 1.0 to be attained within 0.6 s. After the first key frame is attained, the second key frame is executed, which specifies the upper eyelids to move to magnitude 0.3 within 0.3 s. The rendering is done in real time.

To facilitate creating of animations, a graphical editor of facial expressions has been developed by Alzeyara et al. [2011]. The editor visualizes key frames and animations as the designer varies magnitudes of FACS, for either or both sides of the face, as applicable.

Key frames for animations from the experiments are given in Appendix B.1.

6.1.5 Voice

Hala 2's voice is generated by the Acapela text-to-speech software, which provides voices for speaking both American English and Arabic [Acapela Group]. A sound file is generated by the Acapela server and, together with the utterance text, is used to produce *visemes* (animations of phonemes) that are blended with the facial expression in real time [Alzeyara et al., 2011].

6.1.6 Neck

A pan-tilt unit (PTU) is controlled by IM actions that specify the target pan-tilt angles and the attack time. This means that, at present, PTU movement is synchronized with the facial animation only at the IM action level. A future facial expression system will, ideally, include neck positions as part of the key frame specification.

Due to PTU motors' limitations on maximum torque, pan movement has limited acceleration, and tilt movement as a limit on maximum angle from which it can recover the vertical position. The current LCD monitor was found to be too heavy to allow fast tilt at an angle that would make it easily noticable on the video. Therefore, for the video stimuli of the study described in Chapter 4, in addition to a slow physical tilt, we made videos that used an in-screen head tilt instead. Based on the results of that study, we used only in-screen tilt of the head for the final study presented in Chapter 5. However, we used the physical pan accompanied with in-screen pan of the eye gaze in both studies. No in-screen pan movement of the head was used in these studies.

6.2 INTERACTION MANAGER

The interaction manager (IM) is an ongoing implementation of *collaborative discourse* theory of Grosz and Sidner [1990]; Lockbaum [1998]. Its main data structure, a *dialogue tree*, is maintained to represent the current state and is updated depending on the inputs and the recipes, i.e. rules that specify subdialogues. From the interaction perspective, it is designed to enable the following capabilities:

Multi-modal interaction. The user’s utterance and its syntactic/semantic parses are treated in the same fashion as outputs of other sensor processing modules, such as the result of the people-tracking via a laser scanner, or face-tracking via a camera.

Multi-party interaction. Each user has a dedicated data-structure and the interaction recipes can be applied to multiple data structures at once.

Flexible turn taking (mixed initiative). The user may produce an utterance at any time, not necessarily in response to the robot’s prompt (and the utterance will be handled).

Timed execution. Actions can be triggered not only by discrete user actions, but can also by any state variables, including time. For example, the robot may decide to talk to a user after a few seconds of silence.

Flexible discourse structure. Subdialogues, and the nodes of the dialogue tree where the subdialogues are to be applied, are dependent on the current state, including the user input.

Flexible turn interpretation. The user’s utterance can be interpreted in multiple contexts, which are specified by the current dialogue tree.

From the implementation perspective, the M is either a stand-alone process or a C++ library that can be used with various inputs (results of natural language processing subsystems, sensors) and outputs (actuators, natural language and behavior generators). The interaction manager is an open source software that is available via GitHub [Makatchev and Simmons, 2013].

6.2.1 *Related work*

Collaborative agent-based approach to dialogue management is only one of several approaches to dialogue management (see an overview by Bui [2006]). Previously, the definitive implementation of the *Shared-Plan* model [Grosz and Sidner, 1990] of collaborative discourse was the Collagen collaboration manager [Rich and Sidner, 1998], which was proprietary. An open source successor of Collagen, called Disco [Rich and Sidner, 2012], is currently under active development. An extension of SharedPlans that explicitly represents the task structure, called Collaborative Problem-Solving [Blaylock and Allen, 2005], has been used in another proprietary dialogue manager, SAMMIE Becker et al. [2006]. In the development of the IM, we follow the main ideas presented in Collagen and Disco and adapt them for our specific needs.

6.2.2 Overview

Similar to Collagen and Disco, IM creates and maintains a dialogue tree that represents the current state of the dialogue. At present, there is no separate task structure. According to the collaboration discourse theory, a dialogue is viewed as a hierarchy of tasks, and an utterance can contribute to a task in one of the three ways: (1) provide a needed input, (2) select a new task or subtask to work on, or (3) select a recipe to achieve the task. Like Collagen and Disco, IM extends this interpretation paradigm to include inputs of other modalities.

The tree-growing process is a production-like system (c.f. CLIPS expert system [Giarratano and Riley, 1998]) that attempts to satisfy goals by backchaining on rules. All the inputs and state changes are interpreted in the context of the current dialogue tree. According to the joint plan theory of interaction, every participant of the interaction contributes to building the joint plan. Generally, each participant may have their own view of the joint plan, distinct from others. IM considers only one view of the joint plan, which can be considered the robot's point of view.

The knowledge that comes from outside of IM (facts) is represented as logical formulae. In the simple case, these facts form a conjunction of *atoms*, where each atom is a partially grounded predicate.

IM *recipes* have a function similar to Disco and Collagen recipes, and correspond to the rules of a production system. A recipe consists of a precondition, a body and a postcondition. A precondition is a disjunctive normal form (DNF) of partially grounded atoms. A body of the recipe is an ordered sequence of *actions*, *assignments*, or *goals*. The postcondition of a recipe is an assignment.

There are several ways in which IM differs from a typical production system:

- Items in the body of a recipe (corresponding to the right hand side of a production rule) are executed in order (sequentially by default). The execution proceeds to the next item when execution of the previous item is successfully completed.
- There is a mechanism to control the flow of execution depending on the return status of actions. For example, if a time-sensitive action, such as saying “goodbye” to a departing user, has been aborted by the executor, for some reason, the IM may need to purge the rest of the farewell recipe. If that was a greeting, the IM may retry sending it to the behavior executor, or skip it and move to the next action in the greeting recipe.
- There is a mechanism for timing out of items in the body of a recipe.

- There is a whilecondition, failure of which triggers purging of the recipe.

IM's dialogue tree interface implements the three ways of interpretation of an input specified by collaborative discourse theory. Notably, IM allows interleaved recipe execution, resulting in simultaneous multi-topic dialogues. However, at the present state, IM does not implement many of the features of Collagen and Disco, such as focus stack. For more details we refer the reader to the IM documentation at [Makatchev and Simmons, 2013].

6.2.3 *Interfacing with interaction manager*

The interaction manager is intended to hold a central place in the graph of the data flow between the system's components: most actions that are dependent on the sensor inputs should be triggered from the interaction manager. The exceptions are those actions that require low latency (under 100ms) or actions that are dependent on the inputs from one sensor module that maintains its own state. For example, a response of the tiles rendered in the 3D physics simulator to the user's touch of the desktop shared with the Gamebot Victor (see Section 6.2.6) is handled by the simulator's own logic.

The inputs are passed to IM by creating or updating the state atoms (see Section 6.2.4), while IM outputs are actions addressed to particular executors.

A natural language processing pipeline, for example, could feed to IM the surface utterance, its syntactic and semantic parses. IM would produce an action, depending on the input utterance and the current state of the dialogue tree. For instance, given a user's greeting and an empty dialogue tree, the greeting script can be triggered, that could send an action to say "hello" directly to the robot's behavior executor, or could send an action to produce a greeting realization to a natural language generator.

Currently, task-related actions, such as information look-up, are also dispatched via recipes loaded onto the dialogue tree. For example, a recipe handling a weather question first dispatches an action to a weather module for that requests the weather in the particular location and time, and then, based on the returned information, dispatches an action that generates the answer utterance.

Every action specified an excutor that the action is intended to. An executor returns a required action status (completed, aborted, failed) and an optional list of return values. The semantics of action status is flexible, since action call in a recipe allows to override control flow rules conditioned on the action status.

6.2.4 *Interaction manager state*

The complete state of the interaction manager consists of the dialogue tree and the partially grounded atoms corresponding to the system's globals (such as current time and current number of people in the robot's vicinity), user objects (user's last utterance, the time when the user joined, etc.), and other atom types depending on the application. For example, in the IM of a game-playing robot, there would be additional atom types for a player and a game state.

Each node of the dialogue tree, except for the root, corresponds to an instance of a recipe. The nodes are created and purged according to the rules described in Section 6.2.5.

Currently, no analogue of the Collagen or Disco focus stack is implemented. That introduces an undesirable dependency of the IM operation on the order in which recipes are matched against the input.

6.2.5 *Control flow*

At every cycle of its main loop, IM attempts to (a) backchain from the root of the dialogue tree, (b) execute nodes of the dialogue tree, (c) backchain on the pending goals, (d) purge nodes with failed preconditions. We describe these operations in more detail below.

- (a) At every cycle, the preconditions of a subset of the recipes (that are called *triggerable* recipes) are checked against the current state atoms. The preconditions are, in general, a disjunctive normal form made of predicates over the atom slots. The predicates include numeric comparisons for number type slots and regular expression matching for string type slots. For example, a precondition of a recipe for interaction with two interlocutors at the same time can have two user atoms where one of the users has introduced himself (slot name is not equal to `_NO_VALUE_`) and the other has not (slot name equals `_NO_VALUE_`). If such a pair of user atoms exists among the atoms for all the currently present users, it will be matched against the preconditions of the recipe. If multiple matching pairs of atoms exists, one of the pairs (first found by the combinatorial matching algorithm) will be selected. The recipe could then proceed with generating an utterance addressed to the first user "Why don't you introduce your friend?"

Once the formula in the recipe precondition is satisfied, a new node will be created for this recipe and added as a child of the root of the dialogue tree. Local variable bindings specified by the precondition will be created. The first consecutive assignments of the recipe will then be executed to allow the recipe to

set any blocking flags (for example flags that block other recipes from trying to greet the unIntroduced user).

- (b) Once a tree node is created based on a recipe, its body elements are attempted to be sequentially executed, one element per cycle (after the first block of consecutive assignments). If the current element is an action, then the execution pauses until the action status is returned. Then, depending on the returned status and the specified behavior, the execution can move onto the next element (default), skip to the postconditions, or purge the node altogether. If the current element is a goal, then the execution of the node is paused until the goal is satisfied, normally, by backchaining on it with a recipe. There is a way to specify a timeout that will move the execution to the next element if an action or a goal are not completed within the specified time. Once all body elements are executed, the node's postcondition assignment is performed and the node is purged. If the node was created to satisfy a goal of a parent node, that goal now is considered satisfied.
- (c) If a body element is a goal, IM attempts to find a recipe, backchaining on which will satisfy the goal. This is done by matching the assignment in the postcondition of a recipe and the goal DNF. To enable such matching without the need to simulate the execution of the candidate recipe, the postcondition assignment is not allowed to contain variables. Once such a recipe found, a new node is created as a child of the current recipe.

For example, the recipe's goal can be specified as a formula consisting of a single atom for the unIntroduced user, with the `has_been_greeted` slot's value being equal to `true`. Any recipe with the postcondition assigning `true` to the `has_been_greeted` slot of a user's atom can satisfy the goal. However, only a recipe with its preconditions satisfied will be loaded to satisfy the goal. For example, two recipes can greet the user and assign `true` to the `has_been_greeted` slot in the postconditions. One recipe, may be conditioned on the current time of the day being morning, another recipe on the current time being afternoon. Then, out of these two recipes both of which can potentially satisfy the goal, the recipe with the preconditions matching the current time of the day should be selected.

- (d) Finally, at every cycle of the IM loop the whileconditions of every dialogue tree node are checked. Nodes with failed whileconditions are purged, either immediately (in which case executors of pending actions receive a command to abort them), or when the currently pending action receives an action status form its executor.

For example, a dialogue recipe may be conditioned on the user being present in the vicinity of the robot. As soon as the user leaves the vicinity, the corresponding node will be purged, even if it hasn't been fully executed.

6.2.6 Applications

The experiments with the robot prototype in this thesis were Wizard-of-Oz style, so no natural language understanding or sophisticated dialogue trees were necessary. The Wizard-of-Oz control is implemented using goal-free recipes that are triggered by the operator's commands.

IM, with its fuller range of features, is currently used in the following applications.

Victor, the Gamebot, is a Scrabble-playing trash-talking robot installed at the Gates-Hillman Complex of CMU. IM controls Victor's conversation and gameplay strategy, allowing the robot to adjust its level of expertise.

Hala 2 is the next version of Hala, which is currently under development. IM controls the robot's contributions to the conversation.

Uncle Georgi and Fake Siri are radio show sidekicks¹ on the Russian Hour, WRCT Pittsburgh 88.3FM. IM controls the sidekicks' contributions to the conversation.

The Hala 2 character, whose backstory describes it as a bilingual Arabic robot receptionist, should directly benefit from the results of this thesis. The behavioral cues of ethnicity that we used in Chapters 4 and 5 include facial expression animations, neck movements, and different realizations of dialogue acts. The experimental stimuli containing these behaviors have been generated using the Hala 2 software architecture and rendered on the Hala robot prototype. However, engineering such behaviors for a deployment version of a robot receptionist would benefit from the robot's ability to generate their exact realizations in real time, instead of having them hardcoded in the recipes. For example, eye gazing towards the destination at every step of a sequence of directions can be generated on the fly for an arbitrary sequence of directions. Similarly, politeness markers can be injected into an arbitrary direction sequence by the sentence realizer. Such extensions, which would be desirable for a robot that implements results of this study, are discussed as future work in Chapter 7.

¹ [youtube.com/watch?v=B7CRTWY8W4A](https://www.youtube.com/watch?v=B7CRTWY8W4A), [youtube.com/watch?v=1LGCLST8AEQ](https://www.youtube.com/watch?v=1LGCLST8AEQ)

Code	Action unit description
01	Inner brow riser
02	Outer brow raiser
04	Brow lowerer
05	Upper eyelid closer
06	Cheek raiser and lip compressor
07	Eyelid tightener
08	Lips closer
09	Nose wrinkler
10	Upper lip raiser
11	Nosalabial furrow deepener
12	Lip corner puller
13	Sharp lip puller
14	Dimpler
15	Lip corner depressor
16	Lower lip depressor
17	Chin raiser
18	Lip pucker
19	Tongue show
20	Lip stretcher
21	Neck tightener
26	Jaw drop
27	Mouth stretch
28	Lip suck
31	Jaw clencher
32	Bite
33	Blow
38	Nostril dilator
39	Nostril compressor
63.5	Eye X-axis orientation

Table 5: A fragment of codes for action units relevant for the expressions used in this study (based on FACS, Ekman and Friesen [1978]).

SUMMARY AND FUTURE WORK

7.1 SUMMARY OF THE RESULTS

In this thesis, we set out to evaluate two main hypothesis:

Hypothesis I (Ethnic attribution) Believable ethnic identity of robot characters can be created using behaviors selected via lower fidelity on-screen simulations and crowdsourcing.

Hypothesis II (Homophily) Human-robot ethnic congruence improves subjective and objective measures of interactions.

We found partial support for Hypothesis I, and did not find convincing support for Hypothesis II.

We addressed these hypotheses in the context of a robot receptionist (roboceptionist) using Hala hardware and Hala 2 software architectures. In particular, we conducted a sequences of studies that started with identifying candidate behavioral cues of a native speaker of American English (Ar), or a native speaker of Arabic, speaking English as a foreign language (AmE), when viewed by members of either ethnic group. We adopted the notions of a rich point from anthropology and a maximally distinctive behavior from the related work on ethnically salient ECAs, and argued for their distinction from the ethnic cues. This distinction necessitates further evaluation of rich points and maximally distinctive behaviors with respect to their ability to evoke ethnic attribution.

We argued that such an evaluation can be done using low fidelity on-screen stimuli and online crowdsourcing. Our online studies showed that, indeed, perceived ethnic attribution and other perceptual measures of both verbal and nonverbal behaviors can be obtained from both ethnic communities of our focus. However, the obtained scores were scarce and did not give definitive answers by themselves. Nevertheless, equipped with the corpus of human receptionist dialogues, and the intuitions from the related studies on cultural differences, we were able to select candidate cues of ethnicity for their further implementation for the study with a colocated robot prototype.

When expressed by a colocated robot, the behaviors that we identified as likely cues of the AmE and the Ar ethnicities, affected ethnic attributions of the robot character (i.e. combinations of the robot's behaviors and one of the four faces). In particular, the behavioral cues of the AmE ethnicity had a simple main effect on the ethnic attribution

of the characters as AmE by female participants. The behavioral cues of Ar ethnicity had an effect on the Ar attribution for two of the faces.

Notably, ethnic attributions produced by faces rendered on-screen in an online study, were not replicated when the faces were rendered on a robot that expressed behavioral cues of ethnicity, including both the colocated robot and videos of the robot. This could be due to (a) a weaker effect of appearance cues of ethnicity when rendered on a robot, caused by either the perceived machine-like identity of the robot, (b) the perception of a shared physical space, leading to the “broken fourth wall,” or (c) a dominating effect of behavioral cues of ethnicity. A further study could address this issue.

Although we did not find a convincing evidence of ethnic homophily, we believe that suggested pathway can be used to create robot characters with higher degree of perceived similarity, and better chances of evoking homophily effect. We discuss some of the possible ways of improving the perceived similarity in the next section.

Crowdsourcing, while useful, did not resolve all of our problems. For example, we still had difficulties with recruitment of participants from Arabic-speaking countries. Perhaps other online market places could offer a larger fraction of online workers from those countries.

7.2 REFLECTION

Behaviors are an integral part of achieving believability in humans (see, for example, acting techniques of [Stanislavski \[2008\]](#)). As interactive capabilities of robots are gaining in sophistication, robots’ behaviors during these interactions are becoming a major modality for expressing such aspects of believability as emotion, personality, and sociodemographic background, such as ethnicity. Principled data-driven selection of behaviors that are fit to produce the desired effect has not been broadly used yet due to, mostly, sparsity of the data (an exception is the studies by [Mairesse and Walker \[2008\]](#) that we discuss later in this section). In particular, maximally distinctive behaviors between respective groups of users are often assumed to be the salient behaviors that, when implemented on a robot or an on-screen agent, will result in the observer’s attribution of the robot/agent to the corresponding group. This conventional methodology, as applied to the design of ethnically believable characters, is depicted in figure 37.

This approach, while being intuitive, suffers from a number of weaknesses, which may be why its success in achieving believability, in particular, in expressing ethnicity, has been spotty.

First, it does not explicitly take into account interactions between behaviors. For example, it is known that the amount of gaze on the conversation partner depends on who is speaking at the moment. Typically, it is up to the person performing the analysis to account for such dependencies on the context.

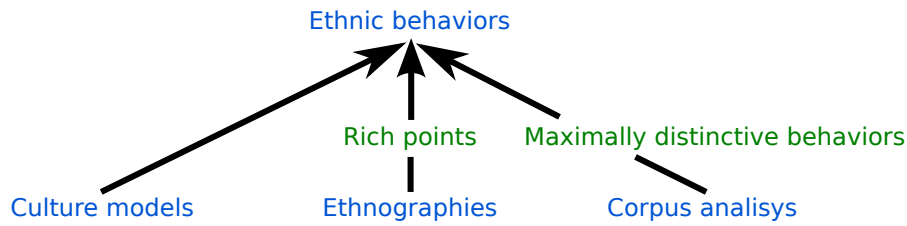


Figure 37: Conventional method for selecting behaviors of ethnic characters.

Second, this approach does not explicitly measure salience of the behaviors for the purposes of the group attribution. Thus, behaviors that are different across the groups because of confound factors, rather than group attribution, will be selected. For example, multimodal corpora usually include relatively small number of subjects, due to the large amount of effort involved in data collection and annotation. These subjects, when divided into the target groups, for example by their ethnicity, will likely not be balanced with respect to other features, such as gender, age, status. These variables may be as important in explaining the differences in behaviors as the ethnicity, but the small number of subjects does allow to control for the effects of these variables.

Third, even when differences are indeed correlated with the target variable, such as ethnicity, their expressive power in eliciting the attribution generally depends on the observer. For example, unless the observer is familiar with the upper head toss as a gesture used to express “no” by some speakers of Arabic, this behavior may not be a strong contributor to eliciting ethnic attribution from this observer. While one may argue that such a behavior at least will not harm attribution, the opposite may be the case: behaviors that are not easily interpretable in the context hurt the believability of human actors [Stanislavski, 2008].

A more sophisticated approach that avoids these problems selects behaviors using procedures of statistical classifier feature selection. Figure 38 illustrates this additional step in application to the design of ethnically believable characters. Such methodology that includes the feature selection step was successfully used by Mairesse and Walker [2008] to generate personality via verbal behaviors but requires considerable amount of training data. A more recent work adds gestures to create on-screen characters that vary their degree of extraversion [Neff et al., 2010]. Adapting this approach to selection of salient ethnic behaviors faces the following problems.

First, it is not clear ethnic attribution of short behavior snippets can be directly evaluated the way the personality can be. The dif-

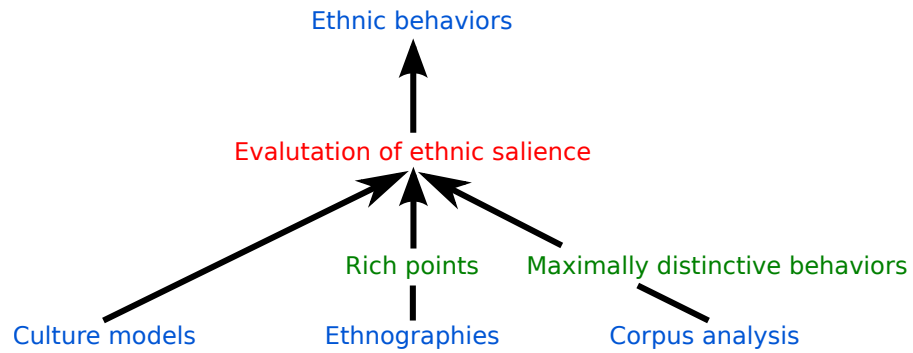


Figure 38: Proposed methodology for selecting behaviors of ethnic characters.

difficulties include lack of established ethnic attribution questionnaires, the need for potentially longer interactions to express ethnicity through behaviors, and some contexts may lend more to the ethnically expressive behaviors than others.

Second, combinations of verbal and nonverbal candidate behavioral cues of ethnicity, when controlled for appearance, create large amount of stimuli that may require special experimental design and recruitment procedures. In particular, it becomes infeasible to test the stimuli on a robot with colocated participants.

Third, as we pointed out, ethnic attribution is likely to be dependent on the ethnicity of the observer, and access to the populations of subjects of target ethnicity can be problematic.

In this thesis we proposed solutions for the problems and verify their feasibility by completing pathway of creating ethnically believable robot characters. We addressed the difficulty of directly evaluating ethnic attribution by using additional measures of perceived animacy, anthropomorphism, likeability, intelligence, and safety (Godspeed questionnaire described by [Bartneck et al. \[2009\]](#)). We addressed the problems of a large number of stimuli and the need to access to participants of particular ethnic background by performing recruitment and evaluation of behaviors via online crowdsourcing.

The substitution of stimuli rendered on a colocated robot by videos of the robot may affect ethnic attribution in multiple ways. It may dampen the effect, as colocated robots were shown by [Kiesler et al. \[2008\]](#) to be more influential and more easily anthropomorphized. It may also strengthen the perceived ethnic attribution, as we speculated that on-screen agents potentially easier than robots lend to the suspension of disbelief as ethnic characters. As we did not separately evaluate the effect of colocation on perceived ethnic attribu-

tion, this remains a topic for a future work. When used jointly in multiple direction-seeking interactions, the behaviors selected based on the evaluation of video stimuli (and additional guidance described below), elicited desired ethnic attribution. However, the methodology uncovered a number of additional problems.

- Few stimuli affected ethnic attribution scores.
- Even when ethnic attribution scores were affected by the stimuli, sometimes the attribution differed across the ethnic groups. Namely, the same verbal behavior, such as saying “Yes, sir” or “Yes, ma’am,” positively affected attribution of the robot as a native speaker of American English when scored by native speakers of Arabic, and negatively affected attribution of the robot as a native speaker of American English when scored by native speakers of American English.
- While using measures of Godspeed questionnaire results in score differences between the ethnic groups, these differences are not as easily interpretable as the ethnic attribution measures. For example, the nonverbal behavior of smiling positively affects all Godspeed measures of native speakers of Arabic, while for native speakers of American English it affects positively perception of animacy, anthropomorphism, likeability, intelligence, but affects negatively the perceived safety.
- Recruitment of the required numbers of participants from Arabic-speaking countries using Amazon’s Mechanical Turk was at times slower than desirable (less than 5 participants per day).

In cases when the interpretation of the ethnic attribution and Godspeed scores was not easily interpretable, we selected the behaviors based on the ethnicity of the original sources of the behavior candidates, or based on the intuitions provided by the ethnographies and culture models. A more systematic handling of such cases would be necessary to completely automate the behavior selection procedure.

As the final stage of the character design, behavioral cues of ethnicity selected via online experiments were implemented on a robot prototype and tested with colocated participants. Although behaviors produced desired affect on attributions (at least for subsets of participants), the absolute attributions of characters were overwhelmingly biased towards native speaker of American English, likely due to the exclusively English content and phonetic characteristics of the text-to-speech system. This suggests using explicit Arabic content (code switching) and prosodic manipulation as additional behaviors that may improve attribution.

In summary, obstacles to the principled data-driven selection of behavioral cues of ethnicity include questionnaire instruments and the

interpretation of the scores they provide, and access to the participants of target demographics. We have shown in this thesis that, in cases when these two obstacles are not insurmountable, the proposed pathway succeeds in creating ethnically believable robot characters.

7.3 FUTURE STUDIES

The current studies had a number of limitations. We did not vary voices, while it is known that prosody, for example, is a rich point [Ward and Bayyari, 2007]. The interaction effects with faces suggest that faces play important, yet complex, roles in the perception of robot characters. In our studies we varied only a limited number of facial features (colors of skin, hair, and eyes). A more representative sample of faces could help shed the light on some of the complex associations between faces and perceptual measures.

More significantly, perhaps, we did not vary the perceived gender of the robot characters. While we controlled for the gender of the participants, varying the gender of the robot character could allow us to add an additional dimension (or two, accounting for the genders of both interlocutors) to the description of rich points and ethnic cues. Of course, new dimensions would increase the amount of work, perhaps exponentially, and would require developing of more advanced experimental machinery. We discuss some of these issues in the next section.

Identifying ethnically salient behaviors turned out to be a difficult task. One possible reason is that native language and the country of residence may play a smaller role in defining a person's social identity than they were just decades ago. However, similarity along other sociodemographic dimensions may be powerful sources of homophily. For example, fans of Pittsburgh Penguins could arguably feel an initial mutual attraction in spite of differences along other dimensions of their social identities. In the limit, a robot may align identity towards an individual interlocutor, effectively personalizing its interaction and increasing the number of dimensions of similarity. However, for a robot receptionist that may need to interact with multiple users at the same time, such personalization to one user may increase a dissonance with others. Finding the appropriate dimensions and degrees of desired similarity between a robot and users in a service scenario can be an interesting direction of future work.

In our studies we have introduced the robot as an actor, playing a character of a receptionist. This was done because we contended that users would more readily apply the notion of ethnicity, or native language, to a character, rather than a robot. We have not, however, evaluated this hypothesis. It is not impossible to imagine that perception of the robot as an actor may reduce homophily effect, even when the ethnic congruence between the user and the robot character

is established. After all, it is not the robot *itself* who is similar to the user, but the *character* the robot plays. Evaluating the effects of agency, that is perceived as either direct or mediated through a character, on ethnic attribution and homophily may be an interesting new line of work.

7.4 IMPLICATIONS FOR TECHNOLOGY

The behavioral cues of ethnicity for the studies described in this thesis were hard coded, rigidly tying verbal and nonverbal modalities. Using such behaviors in a deployment version of the robot would benefit from realtime rendering of the verbal and nonverbal behaviors for an arbitrary, say, direction sequence. Examples of such systems include a model of discourse sensitive posture shifts, developed for Collagen by Cassell et al. [2001], and a natural language generator that expresses personality, developed by Mairesse and Walker [2010].

Identifying and evaluating large numbers of ethnic cues via online studies requires high degree of automation in both generation of stimuli, recruitment of study participants and analysis of the data. With a degree of ingenuity currently lacking in the author, perhaps the problem can be framed as a game with a purpose, alleviating the issues of recruitment and worker remuneration [von Ahn and Dabbish, 2008].

7.5 APPLICATIONS IN SOCIETY

7.5.1 *Cultural competence*

*Everyone should learn basic language skills....
This language skill is as important as
your other basic combat skills.*

— General Stanley A. McChrystal,
Counterinsurgency (COIN) Training Guidance (November, 2009)

In the modern world, cultural awareness is increasingly a matter of survival (cf. Gen. McChrystal, epigraph). Attaining cultural awareness firsthand from members of the languacultures can be infeasible, and even when it is, errors can be costly. Virtual agents and environments have been successfully used to train *cultural competence* (see, for example, [Johnson and Valente, 2008]). While current state of the art in a robot's human-likeness lags behind that of virtual character's, the proposed methodology may be useful in other contexts, besides using a robot as a *nonplayer character* of an educational game. For example, behavioral cues of ethnicity appear to be especially useful for humans, since we have inherent limitations in expressing ethnic cues

via appearance, and attaining a foreign language proficiency may be much harder than learning, for example, a gaze pattern. There is an evidence (see, for example, [Dew and Ward, 1993; Collett, 1971]), that using ethnically congruent nonverbal behaviors improves interaction outcomes even when ethnic attribution is not possible (presumably due to strong appearance cues).

7.5.2 *Facilitating positive intergroup contact*

*I imagine how this robot would be at home, in Palestine:
An Israeli soldier would probably put a gun to its head.*

— A study participant, *A post-study interview*

Homophily has its flip-side: it results in reduced intergroup contact. Overcoming such ingroup communication bias is a worthy goal for a number of reasons. First, positive intergroup communication can reduce prejudice [Nagda, 2006]. Second, communication between members of culturally diverse workgroups is an important factor affecting the workgroup performance [Barinaga, 2007].

One way to reduce ingroup bias is by establishing a contact with an outgroup individual. However, not every intergroup contact has a positive outcome. Allport [1954], in his seminal work, hypothesized that positive effects of intergroup contact require the following four conditions: equal group status within the contact situation, common goals, intergroup cooperation, and the support of authorities, law, or custom. Pettigrew [1998] expanded this list with a friendship potential as an essential fifth condition, and described processes that can contribute to optimal contact.

Everyday situations in many modern societies do not satisfy some or all of the conditions for optimal contact. In Qatar, for example, where nationality is closely linked with occupation, and therefore, status, some combinations of nationalities almost never interact in situations of equal group status [Nagi, 2006]. In other words, if Allport's and Pettigrew's hypotheses are correct, some combinations of ethnic groups in Qatar, particularly involving Qatari nationals, virtually never share situations that are conducive to positive contact.

Perhaps, an ethnically-believable robot character can play roles of an interlocutor or facilitator in a positive intergroup contact situation. Specifically, embedding an ethnically-specific robot character in communicative situations with a member of an outgroup (i.e. a user with an incongruent ethnic identity), may help avoid evoking some of the societal constraints that hinder positive intergroup contact between humans of the corresponding ethnic groups.

APPENDIX

CROSS-CULTURAL PERCEPTION OF PERSONALITY THROUGH LANGUAGE

A.1 STIMULI

*This survey has 602 questions. Really?
Filled to page 7 of 54.*

— A participant in a pilot of this study

Tables 6–10 show stimuli used in the study introduced in Section 4.2. Each of the Tables 7–10 corresponds to a section of the experiment that targets the respective dimension of language variability. Every section of the experiment also contains the control condition shown in Table 6.

Hence, each section of the experiment contained 12 items (excluding one item that was used as a test of diligence of the participants): 4 dialogue turns \times 3 values of valence (including neutral control). Each item was presented on a separate web page. A participant had to complete personality (or naturalness) evaluation to be able to move to the next page (forced choice). Figure 39 depicts the outline of the web page corresponding to one of the items of the study with TIPI personality questionnaire. Similar rendering was used with the version of the study that measured naturalness.

Dimension and valence	Greeting	Question-Answer	Disagreement	Apology
Control (neutral)	A: Good morning. B: Good morning. How may I help you?	A: Could you tell me where the library is? B: It's at the end of the hallway on your left.	A: Could you tell me where the library is? B: It's on the second floor. A: I thought it was on the first floor. B: No, there is no library on the first floor.	A: Could you tell me where the library is? B: Sorry, I don't know.

Table 6: Fragments of dialogues corresponding to the control condition.

Dimension and valence	Greeting	Question-Answer	Disagreement	Apology
Verbosity negative	A: Good morning. B: Morning. May I help you?	A: Could you tell me where the library is? B: End of the hallway on your left.	A: Could you tell me where the library is? B: Second floor. A: I thought it was on the first floor. B: No, it is not.	A: Could you tell me where the library is? B: I don't know.
Verbosity positive	A: Good morning. B: Good morning. How are you today? A: I am doing well, thanks. You? B: Very well, thank you. How's your family? A: Everyone is doing fine, thanks. How about yours? B: Mine is doing well too. How may I help you?	A: Could you tell me where the library is? B: Yes, just follow this hallway until it ends and you will find the library on your left hand side.	A: Could you tell me where the library is? B: Yes, the library is on the second floor. A: I thought it was on the first floor. B: No, there is no library on the first floor of this building.	A: Could you tell me where the library is? B: I am so sorry, I don't really know where the library is in this building.

Table 7: Fragments of dialogues corresponding to the verbosity section of the experiment.

Dimension and valence	Greeting	Question-Answer	Disagreement	Apology
Hedging negative	A: How are you? B: Definitely good, and you? A: Good, thank you. B: How may I help you?	A: Could you tell me where the library is? B: Sure, it is at the end of the hallway on your left.	A: Could you tell me where the library is? B: The library is on the second floor. A: I thought it was on the first floor. B: No, there is definitely no library on the first floor.	A: Could you tell me where the library is? B: Sorry, I have no idea where it is.
Hedging positive	A: How are you? B: Good, I guess, and you? A: Good, thank you. B: Could I help you?	A: Could you tell me where the library is? B: I think it is at the end of the hallway on your left.	A: Could you tell me where the library is? B: I think the library is on the second floor. A: I thought it was on the first floor. B: No, if I am not mistaken, there is no library on the first floor.	A: Could you tell me where the library is? B: Sorry, I am not sure.

Table 8: Fragments of dialogues corresponding to hedging section of the experiment.

Dimension and valence	Greeting	Question-Answer	Disagreement	Apology
Lexical or syntactic alignment negative	A: How are you? B: Good morning. How may I help you?	A: Could you tell me where the bathroom is? B: The restroom is at the end of the hallway on your left.	A: Could you tell me where the bathroom is? B: The restroom is on the second floor. A: I thought the bathroom was on the first floor. B: No, there is no restroom on the first floor.	A: Could you tell me where the bathroom is? B: Sorry, I don't know about the restroom.
Lexical or syntactic alignment positive	A: How are you? B: Good, how are you? How may I help you?	A: Could you tell me where the bathroom is? B: The bathroom is at the end of the hallway on your left.	A: Could you tell me where the bathroom is? B: The bathroom is on the second floor. A: I thought the bathroom was on the first floor. B: No, there is no bathroom on the first floor.	A: Could you tell me where the bathroom is? B: Sorry, I don't know where the bathroom is.

Table 9: Fragments of dialogues corresponding to lexical and syntactic alignment section of the experiment.

Dimension and valence	Greeting	Question-Answer	Disagreement	Apology
Formality negative	A: Good morning. B: What's up? Need anything?	A: Could you tell me where the library is? B: Just go to the end of the hallway, you can't miss it.	A: Could you tell me where the library is? B: Go to the second floor. A: I thought it was on the first floor. B: No, honey, there is none on the first floor.	A: Could you tell me where the library is? B: Sorry about that, I have no idea.
Formality positive	A: Good morning. B: Good morning, sir (madam). Would you allow me to help you with anything?	A: Could you tell me where the library is? B: Kindly follow this hallway and you will encounter the entrance on your left.	A: Could you tell me where the library is? B: Yes, you may find the library on the second floor. A: I thought it was on the first floor. B: I am afraid that is not correct, there is no library on the first floor.	A: Could you tell me where the library is? B: I have to apologize, but I don't know.

Table 10: Fragments of dialogues corresponding to formality section of the experiment.

You	Receptionist
Good morning.	Not so good at all.

Do you agree that the Receptionist's utterance was natural?

Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
1 ○	2 ○	3 ○	4 ○	5 ○	6 ○	7 ○

Table 11: A validation dialogue and question for AmE participants.

You	Receptionist
As-Salamu Alaykum.	Inshallah.

Do you agree that the Receptionist's utterance was natural?

Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
1 ○	2 ○	3 ○	4 ○	5 ○	6 ○	7 ○

Table 12: A validation dialogue and question for Ar participants.

Imagine that you are visiting a university building for the first time and see a female receptionist who appears to be in her early 20s and of the same ethnic background as yourself. The following conversation takes place between you and the female receptionist. (This text will not change from page to page, but the conversations below will be different.)

You	Receptionist
Could you tell me where the library is?	Sorry, I have no idea where it is.

Use the following key to answer the questions below:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly

I see the receptionist as

Extraverted, enthusiastic:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Critical, quarrelsome:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Dependable, self-disciplined:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Anxious, easily upset:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Open to new experiences, complex:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Reserved, quiet:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Sympathetic, warm:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Disorganized, careless:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Calm, emotionally stable:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Conventional, uncreative:

1 <input type="radio"/>	2 <input type="radio"/>	3 <input type="radio"/>	4 <input type="radio"/>	5 <input type="radio"/>	6 <input type="radio"/>	7 <input type="radio"/>
-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------	-------------------------

Figure 39: A rendering of the web page that presents negative hedging of the apology with the personality (TIPI) questionnaire. Symbol ☐ represents HTML radio button element.

A.2 RESULTS: MODEL FITTING

The data set is unbalanced on native language (there were 155 native speakers of Arabic and 166 native speakers of American English) and gender of the participants (163 males and 158 females). Hence, instead of performing a conventional ANOVA with corrections or balancing the dataset by reducing the overall number of participants, we fit the data using linear mixed-effects (LME) models [Pinheiro and Bates, 2000].

LME models are especially suitable to characterize the variation induced by a non-repeatable (for example, study participants) covariate, via so called *random effects*. The covariates with fixed levels, in our case it would be valence, dialogue act, native language and gender covariates correspond to model parameters that we will refer to as *fixed effects*. We divide the data by the language variability dimension and fit LME models separately for each of the subsets, since no comparison across dimensions is intended. Similarly, for the sake of simplicity, we will be fitting models to naturalness score and each of the five TIPI scores individually. In general, it is possible to also fit to a multivariate dependent variable. More specifically, we are considering the following LME models:

1. A random participant factor: $y_{ijk} = \mu + b_i + \epsilon_{ijk}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, and $k = 1, 2, 3, 4$ is a dialogue act index. The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.
2. A fixed valence factor and a random participant factor: $y_{ijk} = \mu + \alpha_j + b_i + \epsilon_{ijk}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, and $k = 1, 2, 3, 4$ is a dialogue act index. The fixed effect of the valence factor is α_j . The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.
3. A fixed dialogue act factor and a random participant factor: $y_{ijk} = \mu + \beta_k + b_i + \epsilon_{ijk}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, and $k = 1, 2, 3, 4$ is a dialogue act index. The fixed effect of the dialogue act factor is β_k . The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.
4. Fixed factors for valence and dialogue act (no interaction) and a random participant factor: $y_{ijk} = \mu + \alpha_j + \beta_k + b_i + \epsilon_{ijk}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, and $k = 1, 2, 3, 4$ is a dialogue act index. The fixed effects of the valence and dialogue act factors are α_j and β_k respectively. The model

This thesis has formulas.

assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

5. Fixed factors for valence and dialogue act (with interaction) and a random participant factor: $y_{ijk} = \mu + \beta_{jk} + b_i + \epsilon_{ijk}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, and $k = 1, 2, 3, 4$ is a dialogue act index. The fixed effects of the valence and dialogue act factors are β_{jk} . The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.
6. Fixed factor for native language not interacting with fixed factors for valence and dialogue act (that interact between each other) and a random participant factor: $y_{ijkl} = \mu + \alpha_l + \beta_{jk} + b_i + \epsilon_{ijkl}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, $k = 1, 2, 3, 4$ is a dialogue act index and $l = 1, 2$ is a native language index. The fixed effect of the native language is α_l and fixed effects of the valence and dialogue act factors are β_{jk} . The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2)$.
7. Fixed factors for native language, valence and dialogue act (all interacting) and a random participant factor: $y_{ijkl} = \mu + \beta_{jkl} + b_i + \epsilon_{ijkl}$, where i is a participant index, $j = 1, 2, 3$ is a valence index, $k = 1, 2, 3, 4$ is a dialogue act index and $l = 1, 2$ is a native language index. The fixed effects of the native language, valence and dialogue act factors are β_{jkl} . The model assumes that the random effect of the study participant $b_i \sim \mathcal{N}(0, \sigma_b^2)$, and error term $\epsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2)$.

Sequential fitting of these and similar models and model selection using ANOVA allows to identify significant effects. Tables 13–16 present the significant effects and the values of χ^2 statistic and p-values for each of the conditions of the experiment. Notice that significant interaction effects between language and valence (and often also dialogue act) covariates are present in all the dimensions of linguistic variability, almost all five TIPI scores and the naturalness score.

Due to the observed significant interaction between valence and act, and three-way interactions between language, valence and dialogue act averaging over dialogue acts is not appropriate.

The selected results of the post-hoc analysis are shown in Figure 25. For a discussion of the results, see Section 4.2.6.

Metric	Effect	χ^2 [d.f.]	p
Extravertedness	valence	[2]113.2	< 0.001
	act	[3]134.5	< 0.001
	valence : act	[6]15.5	0.017
	lang : valence : act	[11]29.5	0.002
Agreeableness	valence	[2]154.3	< 0.001
	act	[3]181.4	< 0.001
	valence : act	[6]35.9	< 0.001
	lang : valence : act	[11]24.4	0.012
Conscientiousness	valence	[2]84.3	< 0.001
	act	[3]328.6	< 0.001
	lang : act	[3]23.1	< 0.001
Stability	valence	[2]36.7	< 0.001
	act	[3]125.8	< 0.001
	gender	[1]3.6	0.058
Openness	valence	[2]104.0	< 0.001
	act	[3]133.5	< 0.001
	lang	[1]10.1	0.002
	gender	[1]4.5	0.062
Naturalness	valence	[2]40.2	< 0.001
	act	[3]101.1	< 0.001
	gender	[1]3.2	0.074
	valence : act	[6]35.1	< 0.001
	lang : valence : act	[11]37.9	< 0.001

Table 13: Significant effects uncovered by model selection on the verbosity subset of the data. Interactions are denoted by colons.

Metric	Effect	χ^2 [d.f.]	p
Extravertedness	valence	[2]34.6	< 0.001
	act	[3]186.6	< 0.001
	valence : act	[6]62.6	< 0.001
	lang : valence : act	[11]23.0	0.017
Agreeableness	valence	[2]4.6	0.098
	act	[3]130.8	< 0.001
	valence : act	[6]96.7	< 0.001
	lang : valence : act	[11]24.0	0.013
Conscientiousness	valence	[2]42.1	< 0.001
	act	[3]208.8	< 0.001
	lang	[1]7.2	0.007
	lang : act	[3]49.5	< 0.001
	lang : valence : act	[11]19.6	0.051
Stability	valence	[2]14.2	< 0.001
	act	[3]67.2	< 0.001
	valence : act	[6]66.1	< 0.001
Openness	valence	[2]25.1	< 0.001
	act	[3]118.7	< 0.001
	lang	[1]5.2	0.023
	valence : act	[6]45.4	< 0.001
	lang : valence : act	[11]41.2	< 0.001
Naturalness	valence	[2]38.9	< 0.001
	act	[3]93.5	< 0.001
	valence : act	[6]27.3	0.001
	lang : valence : act	[11]59.3	< 0.001

Table 14: Significant effects uncovered by model selection on the hedging subset of the data. Interactions are denoted by colons.

Metric	Effect	χ^2 [d.f.]	p
Extravertedness	valence	[2]12.6	0.002
	act	[3]247.9	< 0.001
	lang	[1]4.2	0.041
	valence : act	[6]27.6	0.001
	lang : valence : act	[11]18.7	0.068
Agreeableness	valence	[2]22.8	< 0.001
	act	[3]197.7	< 0.001
	lang : act	[3]21.0	< 0.001
Conscientiousness	act	[3]300.3	< 0.001
	lang	[1]7.2	0.007
Stability	valence	[2]9.7	0.008
	act	[3]147.9	< 0.001
	lang : valence	[2]12.5	0.002
	gender : lang : valence	[2]18.6	0.010
Openness	valence	[2]19.2	< 0.001
	act	[3]127.0	< 0.001
	valence : act	[6]21.7	0.001
Naturalness	valence	[2]73.8	< 0.001
	act	[3]91.7	< 0.001
	lang	[1]6.0	.015
	gender	[1]3.3	0.068
	valence : act	[6]41.7	< 0.001
	lang : valence : act	[11]41.8	< 0.001

Table 15: Significant effects uncovered by model selection on the alignment subset of the data. Interactions are denoted by colons.

Metric	Effect	χ^2 [d.f.]	p
Extravertedness	valence	[2]35.7	< 0.001
	act	[3]152.5	< 0.001
	valence : act	[6]22.9	< 0.001
	lang : valence : act	[11]22.0	0.025
Agreeableness	valence	[2]28.6	< 0.001
	act	[3]38.2	< 0.001
	valence : act	[6]41.4	< 0.001
Conscientiousness	valence	[2]46.3	< 0.001
	act	[3]128.0	< 0.001
	lang	[1]9.8	0.002
	valence : act	[6]29.6	< 0.001
Stability	valence	[2]46.6	< 0.001
	act	[3]25.3	< 0.001
	gender	[1]2.7	0.099
	valence : act	[6]20.4	0.002
	gender : act	[6]11.7	0.008
Openness	valence	[2]24.1	< 0.001
	act	[3]65.9	< 0.001
	lang	[1]20.1	< 0.001
	lang : (valence + act)	[3]12.2	0.007
Naturalness	valence	[2]113.0	< 0.001
	act	[3]30.7	< 0.001
	lang	[1]7.5	.006
	gender	[1]3.3	0.068
	valence : act	[6]74.8	< 0.001
	lang : valence : act	[11]144.4	< 0.001

Table 16: Significant effects uncovered by model selection on the formality subset of the data. Interactions are denoted by colons.

EVALUATING VIDEOS OF VERBAL AND NON-VERBAL BEHAVIORS

B.1 STIMULI

B.1.1 *Questionnaires*

The videos of the stimuli were hosted by Youtube and embedded into a web page that contained a description of the situation and the questionnaires (figures [40](#) and [41](#).) The user could press the “Submit” button at the bottom of the page at any time to proceed to the next stimulus.

Imagine that you are visiting a university building for the first time and the dialog shown in the video happens between you and the robot receptionist. Your speech is shown as the subtitles at the bottom of the video.

تخيل / تخيلي أنك تزور / تزورين مبنى جامعياً للمرة الأولى وأن الحوار الذي يظهر في التسجيل يحصل بينك وبين روبوت الاستقبال، سوف يظهر الجزء الخاص بك من الحوار في أسفل الشاشة



Please rate your impression of the robot on these scales:
صف / صفي انطباعك عن الروبوت على المقياس التالي:

Mechanical	1	2	3	4	5	Organic	فيه انسيابية
Unconscious	1	2	3	4	5	Conscious	مُدرك
Irresponsible	1	2	3	4	5	Responsible	مسؤول
Foolish	1	2	3	4	5	Sensible	مُتزن
Inert	1	2	3	4	5	Interactive	مُتفاعل
Unfriendly	1	2	3	4	5	Friendly	ودود
Awful	1	2	3	4	5	Nice	لطيف
Ignorant	1	2	3	4	5	Knowledgeable	مُطلع
Dislike	1	2	3	4	5	Like	أحبّه
Moving rigidly	1	2	3	4	5	Moving elegantly	حركته أنيقة
Machinelike	1	2	3	4	5	Humanlike	شبيه بالإنسان
Incompetent	1	2	3	4	5	Competent	كفوء
Unpleasant	1	2	3	4	5	Pleasant	مُمتع

Figure 40: A single stimulus web page presents the description of the situation, the video of the stimulus, items of the Godspeed questionnaire with their Arabic translations and the ethnic attribution items. Continued in figure 41.

Unintelligent	1	2	3	4	5	Intelligent
غير ذكي						ذكي
Fake	1	2	3	4	5	Natural
مُفَوِّف						طبيعي
Artificial	1	2	3	4	5	Lifelike
مصطنع						حقيقي
Unkind	1	2	3	4	5	Kind
لئيم						طيِّب
Dead	1	2	3	4	5	Alive
ميت						حي
Stagnant	1	2	3	4	5	Lively
سكن						حيوي
Apathetic	1	2	3	4	5	Responsive
لئيم						مُجاوب

Please rate your emotional state on these scales:

صِف / صِفِي حَالَتَكَ العاطفية على السلم التالي:

Anxious	1	2	3	4	5	Relaxed
مضطرب / مضطربة						مُبتَرِّح / مُبتَرِّحة
Agitated	1	2	3	4	5	Calm
مضطرب / مضطربة						مطمئن / مطمئنة
Quiescent	1	2	3	4	5	Surprised
هادئ / هادئة						مُفاجئ / مُفاجئة


The robot's utterances and movement copy the utterances and movements of a real person. Please rate the likely native language of that person:

تَحْكِي في قول الروبوت بحركاته في قول بحركات شخص حقيقي بربك، ما هي اللغة الأم له هذا الشخص:

Not American English	1	2	3	4	5	American English
لغس اللغة الانكليزية الأمريكية						اللغة الانكليزية الأمريكية
Not Arabic	1	2	3	4	5	Arabic
لغس اللغة العربية						اللغة العربية

Figure 41: Continued from figure 40. A single stimulus web page presents the description of the situation, the video of the stimulus, items of the Godspeed questionnaire with their Arabic translations and the ethnic attribution items.

t2



What did the robot say?

Hello	<input type="radio"/>
Goodbye	<input type="radio"/>
Good evening	<input type="radio"/>
How are you?	<input type="radio"/>

Figure 42: A validation question for AmE participants.



Figure 43: A validation question for Ar participants. It asks “What did the robot say?” and provides a number of choices, only one of which is correct.

B.1.2 Greeting

The greeting varied with respect to face (Face1, Face2, Face3), verbal action (“Hi,” “Yes, ma’am,” or “Yes, sir”), and nonverbal action (no action, open smile, physical nod, or in-screen nod). One of the greetings “Yes, ma’am” or “Yes, sir” was chosen for each participant, depending on their self-reported gender. All greeting nonverbal behaviors start with a neutral face and the robot’s eyes are initially gazing downward (refer to Section 6.1.3 for an introduction to the key frame syntax):

```
(
  (
    ('01', .0, .1), ('02', .0, .1), ('04', .0, .1),
    ('05', 0.1, .1), ('06', .0, .1), ('07', .4, .1),
    ('08', .0, .1), ('09', .0, .1), ('10', .0, .1),
    ('11', .0, .1), ('12', .0, .1), ('13', .0, .1),
    ('14', .0, .1), ('15', .0, .1), ('16', .0, .1),
    ('17', .0, .1), ('18', .0, .1), ('19', .0, .1),
    ('20', .0, .1), ('21', .0, .1), ('26', .0, .1),
    ('27', .0, .1), ('28', .0, .1), ('31', .0, .1),
    ('32', .0, .1), ('33', .0, .1), ('38', .0, .1),
    ('39', .0, .1), ('63.5', -0.2, 0.1)
  )
)
```

At 0.5s the caption of the assumed user utterance appears. At the 2s mark, the caption disappears. After the 2s mark, the nonverbal behaviors differ across conditions as follows.

No action. At the 2s mark the robot’s eyes start moving upwards, to a forward horizontal gaze, taking 0.1s to complete the movement. As the eyes move towards their neutral position and settle there, the robot speaks one of the greeting utterances. The video ends 2s after the robot stops speaking.

Open smile. At the 2s mark, as the robot speaks its utterance, the following smile expression is executed:

```
(
  (
    ('01', 0.5, 0.3), ('05', 0.45, 0.1), ('06', 1.0, 0.6),
    ('12', 0.6, 0.6), ('13', 0.4, 0.6), ('63.5', .0, .1)
  )
)
```

After the robot finishes speaking, it drops the eyebrows and eyelids to their neutral position, while keeping the smile:

```
(
  (
    ('10', 0.3, 0.6), ('11', 0.3, 0.1), ('16', 0.2, 0.6)
  ),
  (
    ('01', 0, 0.3), ('05', 0.3, 0.3)
  )
)
```

The video ends 2s after the robot stops speaking.

Physical nod. At the 2s mark, as the robot's eyes start moving upwards and the robot starts speaking, the neck tilts downward 0.12rad in 2s and then back to the horizontal position in 3s. The slower upward move of the tilt motor is necessary due to the limits on maximum torque.

In-screen nod. At the 2s mark, as the robot starts speaking, the robot's 3d head model was manually controlled (due to the lack of an API for this feature in the version of the CONCEPT software available at the time) to tilt downwards to about 0.5rad within about 0.5s and back up to its horizontal position within about 0.5s.

The URLs for all the stimuli of the greeting dialogue act are shown below.

"Hi" + open smile:

hb1: <http://youtu.be/8bswX-EQf08>

hb2: <http://youtu.be/olzqLihmNSU>

hb3: http://youtu.be/Fvc_43h8AUc

"Yes, ma'am" + physical nod:

ym1: <http://youtu.be/S6Ncj65JH6c>

ym2: <http://youtu.be/4S869a6pcKs>

ym3: <http://youtu.be/MdcuNzKi2d8>

"Yes, ma'am" + in-screen nod:

ymn1: <http://youtu.be/Cev300jCHrc>

ymn2: http://youtu.be/GZA8k_2YHAo

ymn3: <http://youtu.be/b7VJVV9wGlS>

"Yes, sir" + physical nod:

ys1: <http://youtu.be/tmdZGnkC04w>

ys2: http://youtu.be/Dpnm3S0X_w

ys3: http://youtu.be/_DdkLTkuKYA

"Yes, sir" + in-screen nod:

ysn1: <http://youtu.be/gWvjh6a0SEU>

ysn2: <http://youtu.be/wknV-sTWHak>
 ysn3: <http://youtu.be/ZnhPl7bhbtI>

“Hi”:

hnu1: <http://youtu.be/QP1yKXVx18I>
 hnu2: <http://youtu.be/z2DqES7Grmc>
 hnu3: <http://youtu.be/2KBNAvuPi4U>

“Hi” + physical nod:

hrn1: <http://youtu.be/ttKRsj7frF4>
 hrn2: <http://youtu.be/bM8obHI8jWc>
 hrn3: <http://youtu.be/uDjsvSUDMFE>

“Hi” + in-screen nod:

hvn1: <http://youtu.be/jfrG1wFgBLs>
 hvn2: <http://youtu.be/HM2F2rCkHk8>
 hvn3: <http://youtu.be/vjEEY4cHEbE>

“Yes, ma’am”:

ymnu1: http://youtu.be/jBBv_1q7qdQ
 ymnu2: <http://youtu.be/hkcpXpCl5v0>
 ymnu3: <http://youtu.be/etnKdZEU8NU>

“Yes, ma’am” + open smile:

yms1: <http://youtu.be/ZAIr9Aj8S8A>
 yms2: <http://youtu.be/LmvQvpQZu08>
 yms3: <http://youtu.be/haAKYOKnu6g>

“Yes, sir”:

ysnu1: <http://youtu.be/hEUSpp1BlVE>
 ysnu2: <http://youtu.be/kX-SYUG701g>
 ysnu3: <http://youtu.be/XUSa3tE-0DU>

“Yes, sir” + open smile:

yss1: http://youtu.be/lSqhj0_yIZI
 yss2: <http://youtu.be/FmqplKMcY14>
 yss3: http://youtu.be/DmmWn_3vwhI

B.1.3 *Direction giving: Gaze*

The direction giving gaze stimuli had 6 conditions based on the amount of total gaze and the durations of the gaze intervals. All conditions started with the caption of user’s question “Can you tell me where the library is?” appearing from 0.5s to 2s marks of the video, while

the robot's face is in its neutral position. The robot's nonverbal behavior after the first 2s varied according to conditions.

Gaze1. As the robot begins answering with the directions to the library, its physical neck pans 0.4rad towards the robot's right within 0.3s. At the same time, the robot's eyes gaze to the right, executing the following animation:

```
(
  (
    ('61.5', -0.2, .3),
  )
)
```

When all four steps of the directions are given, specifically, 1.2s after it starts giving the last step "Turn left at the hallways and you will see the library doors," the robot's physical neck and eyes pan back into forward position over the duration of 0.3s.

lib11: <http://youtu.be/JhsyGmwyHVw>

lib12: http://youtu.be/nZkTfRJ_rDg

lib13: <http://youtu.be/bQV9QbmB73k>

Gaze2. In this condition, the robot attains the rightward gaze (with the corresponding neck and eye pan) over the first 0.3s of each dialogue turn and starts panning its gaze back to its neutral position 1.2s after beginning of each turn. As the robot begins answering with the directions to the library, its physical neck pans 0.4rad towards the robot's right and then back to its neutral position, returning to its forward direction about 1.5s after the beginning of the robot's turn—at about the same time as its eyes attain their neutral position as well. Thus, according to our convention on measuring the gaze duration as the interval between two onsets, the duration of such pointing gaze is 1.2s. After 0.1s pause, the next turn begins.

lib21: <http://youtu.be/XaQGgbhBDy8>

lib22: <http://youtu.be/ECvHWBwBx9s>

lib23: <http://youtu.be/8HggEofUjIM>

Gaze3. Same as Gaze2, but the robot gazes to the right only during first and third turns.

lib31: <http://youtu.be/qBaZtxGvV74>

lib32: http://youtu.be/e9vG_rAV_2Y

lib33: <http://youtu.be/Md2IvwQAuSM>

Gaze4. Same as Gaze2, but the robot gaze to the right lasts 0.8s, instead of 1.2s.

lib41: <http://youtu.be/L9XLWPHphuk>

lib42: <http://youtu.be/K9QF4ExjE0M>

lib43: http://youtu.be/mnMd_0ApvSI

Gaze5. Same as Gaze4, but the robot gazes to the right only during first and third turns.

lib51: <http://youtu.be/LM-SWtmJD3E>

lib52: <http://youtu.be/SgM60rJRvnA>

lib53: <http://youtu.be/oKIEs3QJirQ>

Gaze6. The robot's gaze remains in the neutral position over the whole duration of the dialogue.

lib61: <http://youtu.be/aP6tj9pfaAE>

lib62: <http://youtu.be/5rZH5C0YQpE>

lib63: <http://youtu.be/HjsJjcmLDQA>

B.1.4 *Handling failure to provide an answer*

No idea. In the first condition, the robot's face displays the animation that includes lip stretcher:

```
(
  (
    ('16', 0.6, 0.6), ('20', 0.4, 0.6), ('13', 0.3, 0.6),
    ('01', 0.5, 0.3)
  )
)
```

It then transitions its gaze to the left over the duration of 0.6s, including a 0.2rad physical neck pan. As the robot starts returning its gaze to the neutral direction over the next 2s, it says "I have absolutely no idea."

noidea1: <http://youtu.be/6wA05CdFGUU>

noidea2: <http://youtu.be/Sv1T3D9pENU>

noidea3: <http://youtu.be/Rvax6NHlgcE>

Explanation. In the second condition, the robot transitions the gaze left and keeps it there for the total of 2.3 seconds:

```
(
  (
    ('61.5', 0.2, .5)
  ),
  (
    ('61.5', 0.2, 1.8)
  )
)
```

including the neck pan of 0.2rad, as it says “...I don’t know where it is...” Then, as the robot starts returning the gaze to its neutral position over the next 1s, it says “...because I am new...”

dontknow1: <http://youtu.be/sYe-iQL9CG0>

dontknow2: http://youtu.be/xhoZ6V_2MZE

dontknow3: <http://youtu.be/yIzwFLA6z9U>

B.1.5 *Handling disagreement*

The robot’s gaze remains in the neutral position over the whole duration of this dialogue.

yesright1: http://youtu.be/RqiXkVUc_hg

yesright2: http://youtu.be/p_5PMPwfttg

yesright3: <http://youtu.be/b0ev2jRYa58>

noright1: <http://youtu.be/3z0a3MN5FuI>

noright2: <http://youtu.be/3pawS-zv-4w>

noright3: http://youtu.be/ogGiD__nGHA

B.1.6 *Direction giving: Politeness*

The nonverbal behavior in these two conditions is equivalent to Gaze3 described in Section B.1.3.

lib31: <http://youtu.be/qBaZtxGvV74>

lib32: http://youtu.be/e9vG_rAV_2Y

lib33: <http://youtu.be/Md2IvwQAuSM>

lib3p1: <http://youtu.be/0YLCpVGg1kU>

lib3p2: <http://youtu.be/hpXdgNVRpkg>

lib3p3: <http://youtu.be/PSJyFMetAa8>

B.1.7 *User validation questions*

These are the two stimuli used to validate diligence of the crowd-sourced worker. The first stimuli was used with AmE participants, the second was for Ar participants. We asked the workers to select the robot’s utterance out of 4 possible choices.

good evening (t1): <http://youtu.be/JrxJrpgS6M8>

marhaba (t2): <http://youtu.be/dwsrJV7H6Is>

B.2 RESULTS

Tables 17–26 show signs of 95% HPD confidence intervals for the term coefficients of Godspeed measures and attribution, when the intervals do not include zero. We abbreviate animacy as ani, antropomorphis as ant, likeability as l, intelligence as i, and safety as s. Attribution measures are, as usual, AmE and Ar.

Face	Verbal	Nonverbal	AmE	Ar
1	"Yes, sir" or "Yes, ma'am"	physical nod		
2	"Yes, sir" or "Yes, ma'am"	physical nod		
3	"Yes, sir" or "Yes, ma'am"	physical nod		
1	"Yes, sir" or "Yes, ma'am"	in-screen nod	l+ ani+	i+ l+ ani+ ant+ s+
2	"Yes, sir" or "Yes, ma'am"	in-screen nod	l+	ant+
3	"Yes, sir" or "Yes, ma'am"	in-screen nod		ani+ ant+ s+
1	"Yes, sir" or "Yes, ma'am"	neutral		
2	"Yes, sir" or "Yes, ma'am"	neutral		
3	"Yes, sir" or "Yes, ma'am"	neutral		ani-
1	"Yes, sir" or "Yes, ma'am"	smile	l+	i+ l+ ani+ ant+ s+
2	"Yes, sir" or "Yes, ma'am"	smile		l+ ani+ ant+
3	"Yes, sir" or "Yes, ma'am"	smile		i+ l+ ani+ ant+
1	"Hi"	smile	l+	i+ l+ ani+ ant+ s+
2	"Hi"	smile		i+ l+ ani+ ant+ s+
3	"Hi"	smile		i+ l+ ani+ ant+ s+
1	"Hi"	neutral		ani-
2	"Hi"	neutral	ani-	AmE- ani-
3	"Hi"	neutral	ani- ant-	ani-
1	"Hi"	physical nod		
2	"Hi"	physical nod		
3	"Hi"	physical nod		
1	"Hi"	in-screen nod		i+ ant+
2	"Hi"	in-screen nod		
3	"Hi"	in-screen nod		

Table 17: Perception of greetings. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero.

Term of a linear mixed effects model	AmE	Ar
face 2		s-
face 3	ani- ant- s-	l- ani- ant-
"Yes, sir" / "Yes, ma'am"	AmE- i+ l+ ani+ ant+	AmE+
physical nod	ani+ s-	ani+ ant+
in-screen nod	i+ l+ ani+ ant+	i+ l+ ani+ ant+ s+
smile	i+ l+ ani+ ant+ s-	i+ l+ ani+ ant+ s+
face 2 : "Yes, sir" / "Yes, ma'am" face 3 : "Yes, sir" / "Yes, ma'am"		
face 2 : physical nod face 3 : physical nod face 2 : in-screen nod face 3 : in-screen nod face 2 : smile face 3 : smile		s-
"Yes, sir" / "Yes, ma'am" : physical nod "Yes, sir" / "Yes, ma'am" : in-screen nod "Yes, sir" / "Yes, ma'am" : smile		i- l- s-

Table 18: Significant terms (up to two-way interactions) for scores of greeting stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1, neutral facial expression, and "Hi" are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero.

Face	Verbal	Nonverbal (point. gaze)	AmE	Ar
1	directions	$1 \times 10s$		
2	directions	$1 \times 10s$		
3	directions	$1 \times 10s$	s-	l-
1	directions	$4 \times 1.2s$		
2	directions	$4 \times 1.2s$	s-	
3	directions	$4 \times 1.2s$	l-	l-
1	directions	$2 \times 1.2s$		
2	directions	$2 \times 1.2s$		
3	directions	$2 \times 1.2s$		l- ant-
1	directions	$4 \times 0.8s$		
2	directions	$4 \times 0.8s$	Ar-	
3	directions	$4 \times 0.8s$	l- s-	l- s-
1	directions	$2 \times 0.8s$		
2	directions	$2 \times 0.8s$		
3	directions	$2 \times 0.8s$		
1	directions	none	AmE+ Ar- l- ani-	l- ani- ant-
2	directions	none	l- ani- s-	i- l- ani- ant- s-
3	directions	none	i- l- ani- ant- s-	l- ani- ant- s-

Table 19: Perception of pointing gaze during direction giving. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero.

Term of a mixed effects model	AmE	Ar
face 2 face 3	AmE- l- ani- ant- s-	Ar- l- ani- ant- s-
gaze 2 gaze 3 gaze 4 gaze 5 gaze 6	 Ar- AmE+ i- l- ani- ant-	ani+ ani+ i- l- ani- ant- s-
face 2 : gaze 2 face 3 : gaze 2 face 2 : gaze 3 face 3 : gaze 3 face 2 : gaze 4 face 3 : gaze 4 face 2 : gaze 5 face 3 : gaze 5 face 2 : gaze 6 face 3 : gaze 6	 AmE-	

Table 20: Significant terms (up to two-way interactions) for scores of direction-giving stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and gaze 1 are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero.

Face	Verbal	Nonverbal	AmE	Ar
1	no explanation	lip stretch		
2	no explanation	lip stretch		
3	no explanation	lip stretch		l-
1	explanation	no lip stretch	s-	
2	explanation	no lip stretch		
3	explanation	no lip stretch		ant-

Table 21: Perception of failures to provide an answer. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero.

Term of a mixed effects model	AmE	Ar
face 2		
face 3		l- ant-
no explanation + lip stretch	s+	Ar- AmE+
face 2 : no explanation + lip stretch	s-	
face 3 : no explanation + lip stretch		

Table 22: Significant terms (up to two-way interactions) for scores of failure stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and explanation are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero.

Face	Verbal	Nonverbal (point. gaze)	AmE	Ar
1	implicit	$1 \times 1.4s$		
2	implicit	$1 \times 1.4s$		
3	implicit	$1 \times 1.4s$		
1	explicit	$1 \times 1.4s$	i+	
2	explicit	$1 \times 1.4s$	i+	
3	explicit	$1 \times 1.4s$	i+	l-

Table 23: Perception of disagreements. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero.

Term of a mixed effects model	AmE	Ar
face 2		
face 3		l- ant- s-
implicit	i- l+ ani-	i- l+
face 2 : implicit		
face 3 : implicit		

Table 24: Significant terms (up to two-way interactions) for scores of disagreement stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and explicit disagreement are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero.

Face	Verbal	Nonverbal (point. gaze)	AmE	Ar
1	direct	$2 \times 1.2s$		
2	direct	$2 \times 1.2s$		
3	direct	$2 \times 1.2s$	l- ani-	l- ani-
1	polite	$2 \times 1.2s$		
2	polite	$2 \times 1.2s$		
3	polite	$2 \times 1.2s$	i-	

Table 25: Perception of politeness markers. + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not include zero.

Term of a mixed effects model	AmE	Ar
face 2		
face 3	i-	l- ant-
polite	l+	i+ l+ ani+
face 2 : polite		
face 3 : polite		

Table 26: Significant terms (up to two-way interactions) for scores of politeness stimuli (intercepts excluded) by AmE and Ar participants (divided by columns). Face 1 and direct style are the reference values for each of the independent variables and hence are not shown. Labels + and - denote signs of 95% HPD confidence intervals for the term coefficients, when the intervals do not contain zero.

EVALUATING ROBOT BEHAVIORS

C.1 STIMULI

C.1.1 *Protocol*

Participants were verbally reminded of the interaction protocol for their next task before each of the three tasks with each of the four robot characters. Additionally, a printed copy of the protocol, shown in figures 44–46, was handed to the participant. The destination in the first task varied between the offices of Prof. Adams, Brown, Coopers, or Douglas.

In this dialogue you will ask for directions to Professor Adams office. Later you may be asked to recall this directions.
Please follow these instructions:

- Greet the receptionist
- Ask the receptionist for directions to Professors Adams' office.
- Say something to end the conversation.

Figure 44: Task 1 interaction protocol that was given to participants.

In this task you will ask the receptionist for directions to Cafeteria. Please follow these instructions:

- Greet the receptionist
- Ask the receptionist for directions to Cafeteria.
- Pretend that you did not understand clearly and ask if you should turn left after the door.
- Say something to end the conversation.

Figure 45: Task 2 interaction protocol that was given to participants.

In this task you will ask the receptionist for directions to the Dean's office .

Please follow these instructions:

- Greet the receptionist
- Ask the receptionist for directions to the Dean's office.
- Say something to end the conversation.

Figure 46: Task 3 interaction protocol that was given to participants.

C.1.2 Questionnaires

The demographic (figure 47) and baseline emotional score (figure 48) questionnaires were given to the participants before they were introduced to any of the robot characters.

Demographic information

Please provide us with the following information about yourself. This information (as well as the rest of the experiment) will remain completely anonymous. Please answer as many questions as possible.

Age:

Sex: male female

Native language:

Please specify your race:
 American Indian or Alaska Native
 Asian
 Black or African American
 Native Hawaiian or Pacific Islander
 White

Are you Hispanic/Latino?

Please rate your English proficiency:

	Beginner level	Not very good	Adequate	Very good	Native speaker level
Reading					
Writing					
Speaking					
Understanding spoken language					

List 3 countries in which you have lived the longest:

Country	Number of years you lived there

Figure 47: The demographic questionnaire.

Please rate your emotional state on these scales:
 صف / صف في حالتك الانعاطفية على السلم التالي:

Anxious	1	2	3	4	5	Relaxed
مضطرب / مضطربة						مسترخي / مسترخية
Agitated	1	2	3	4	5	Calm
مضطرب / مضطربة						مطمئن / مطمئنة
Quiescent	1	2	3	4	5	Surprised
هادئ / هادئة						مفاجئ / مفاجئة

Figure 48: The baseline emotional state questionnaire.

Please rate your impression of the robot on these scales:
 صيف / صيفي انطباعك عن الروبوت على المقياس التالي:

Mechanical آلي	1	2	3	4	5	Organic فيه انسيابية
Unconscious غير مُدرك	1	2	3	4	5	Conscious مُدرك
Irresponsible عديم المسؤولية	1	2	3	4	5	Responsible مسؤول
Foolish أخرق	1	2	3	4	5	Sensible مُثزن
Inert جامد	1	2	3	4	5	Interactive مُتفاعل
Unfriendly غير ودود	1	2	3	4	5	Friendly ودود
Awful مقيت	1	2	3	4	5	Nice لطيف
Ingnorant جاهل	1	2	3	4	5	Knowledgeable مُطلع
Dislike أكرهه	1	2	3	4	5	Like أحبه
Moving rigidly حركته متصلبة	1	2	3	4	5	Moving elegantly حركته أليفة
Machinelike شبيه بالآلات	1	2	3	4	5	Humanlike شبيه بالإنسان
Incompetent غير كفوء	1	2	3	4	5	Competent كفوء
Unpleasant تغيض	1	2	3	4	5	Pleasant مُمتع
Unintelligent غير ذكي	1	2	3	4	5	Intelligent ذكي
Fake مزيف	1	2	3	4	5	Natural طبيعي
Artificial اصطناعي	1	2	3	4	5	Lifelike حقيقي
Unkind لئيم	1	2	3	4	5	Kind طيب
Dead يبدو ميتاً	1	2	3	4	5	Alive يبدو حياً
Stagnant ساكن	1	2	3	4	5	Lively حيوي
Apathetic لا مبالى	1	2	3	4	5	Responsive مُتجاوب

Please rate your emotional state on these scales:
 صيف / صيفي حالتك العاطفية على المقياس التالي:

Anxious متوتر / متوترة	1	2	3	4	5	Relaxed مُسترخي / مُسترخية
Agitated مضطرب / مضطربة	1	2	3	4	5	Calm مطمئن / مطمئنة
Quiescent هادئ / هادئة	1	2	3	4	5	Surprised متفاجئ / متفاجئة

Figure 49: Page 1 of the questionnaire: items of the Godspeed questionnaire with their Arabic translations.

This robot acts as a receptionist character. We ask you guess that character's ethnicity, in terms of the character's native language. Please rate your guess:

يُقال هذا الريبوت دور موظف استقبال. نودّ منك أن تخمّن ethnicity هذا الموظف بحسب بلده الأم:

Not American English	1	2	3	4	5	American English
لست اللغة الناطقة الإنجليزية						اللغة الناطقة الإنجليزية
Not Arabic	1	2	3	4	5	Arabic
لست اللغة العربية						اللغة العربية

Please write down the directions to the Professor's office, to the best of your recollection:

Figure 50: Page 2 of the questionnaire: ethnic attribution items with their Arabic translations, and the written part of map task.

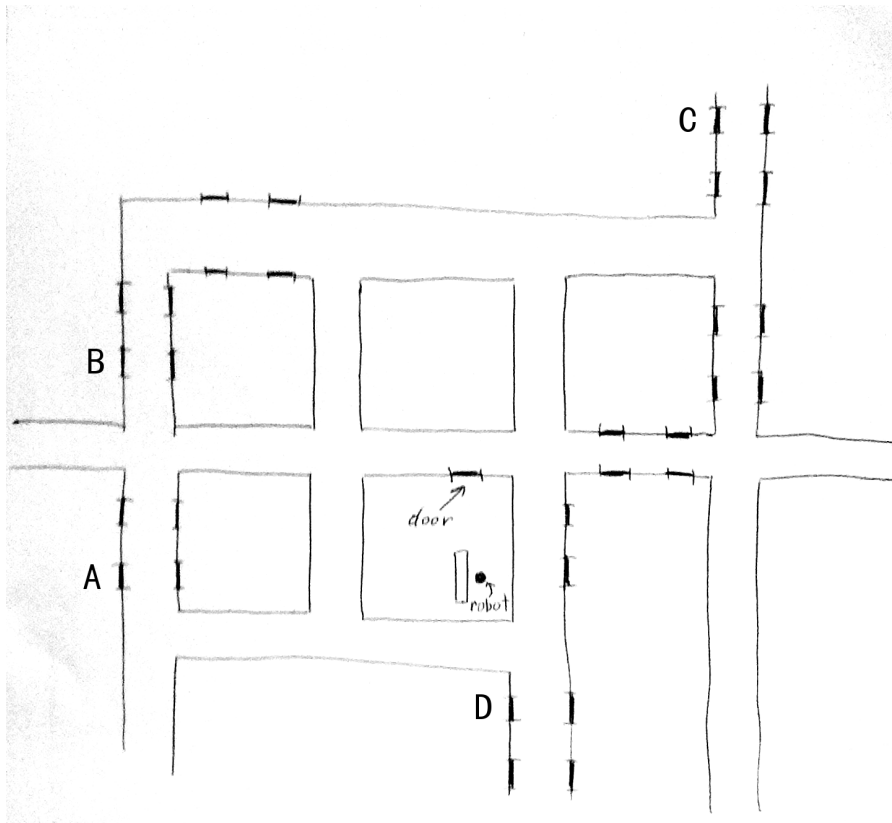
C.1.3 *Locating the destination on the map*

Figure 51: The participants were given this map and asked to draw directions and mark the destination office. The destination labels (A for Adams, B for Brown, C for Coopers, and D for Douglas) were not shown to the participants. The location of the robot and the door to the lab where the experiment was conducted were included for reference.

C.2 RESULTS

C.2.1 Attribution of the robot characters as Ar

Term	Estimate	t value	Pr(> t)
Intercept	1.730	18.578	< 0.0002
Face3 : male	0.966	1.796	0.076
Face4 : BCE2	-1.760	-2.251	0.027

Table 27: Significant associations with the robot characters' attribution as Ar by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	2.426	3.903	< 0.0004
YouAreWelcome	-0.676	-1.700	0.097
Face3 : male	2.919	2.486	0.017
BCE2 : Face3 : male	-3.0543	-1.982	0.055

Table 28: Significant associations with the robot characters' attribution as Ar by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	1.240	2.906	0.006
YouAreWelcome	0.640	2.275	0.028
EC2 : Face2	-1.400	-1.866	0.069
EC2 : Face4	-1.642	-2.172	0.036

Table 29: Significant associations with the robot characters' attribution as Ar by Ar participants. Interactions are denoted by colons.

c.2.2 Attribution of the robot characters as AmE

Term	Estimate	t value	Pr(> t)
Intercept	3.531	18.698	< 0.0001
BCE2	0.688	2.574	0.011
BCE2 : male	-0.652	-1.667	0.098

Table 30: Significant associations with the robot characters' attribution as AmE by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	4.000	8.803	< 0.0001
Face2	-1.200	-1.867	0.068
Face3 : male	-2.267	-2.009	0.050
BCE2 : Face3 : male	3.217	2.193	0.033

Table 31: Significant associations with the robot characters' attribution as AmE by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	3.789	25.05	< 0.0001

Table 32: Significant associations with the robot characters' attribution as AmE by Ar participants.

C.2.3 Animacy of robot characters

Term	Estimate	t value	Pr(> t)
Intercept	3.472	27.952	< 0.0001
BCE2	0.244	2.142	0.034
YouAreWelcome	-0.448	-3.474	< 0.0008

Table 33: Significant associations with the animacy scores of robot characters by all participants.

Term	Estimate	t value	Pr(> t)
Intercept	4.065	12.359	< 0.0001
Face4	-0.811	-2.428	0.019
YouAreWelcome	-0.725	-3.172	0.003
BCE2 : Face4	0.822	1.989	0.052
Face2 : male	0.729	1.778	0.081
Face4 : male	0.718	1.731	0.089

Table 34: Significant associations with the animacy scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	3.533	26.896	< 0.0001

Table 35: Significant associations with the animacy scores of robot characters by Ar participants.

c.2.4 *Anthropomorphism of robot characters*

Term	Estimate	t value	Pr(> t)
Intercept	3.531	24.727	< 0.0001
AmE	-0.353	-2.277	0.025
YouAreWelcome	-0.294	-1.692	0.093

Table 36: Significant associations with the anthropomorphism scores of robot characters by all participants.

Term	Estimate	t value	Pr(> t)
Intercept	3.403	10.340	< 0.0001
Face4	-0.793	-1.919	0.060
male	-1.016	-2.839	0.006
LocationD	-0.504	-1.969	0.054
BCE2 : Face4	0.943	1.849	0.070
Face2 : male	1.402	2.757	0.008
Face3 : male	1.093	2.119	0.039
Face4 : male	1.308	2.546	0.014

Table 37: Significant associations with the anthropomorphism scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	3.540	22.650	< 0.0001

Table 38: Significant associations with the anthropomorphism scores of robot characters by Ar participants. Interactions are denoted by asterisks.

C.2.5 *Likeability of robot characters*

Term	Estimate	t value	Pr(> t)
Intercept	4.212	18.610	< 0.0001
YouAreWelcome	-0.547	-3.653	< 0.0004
BCE2 : Face4	0.829	2.273	0.025

Table 39: Significant associations with the likeability scores of robot characters by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	4.083	11.410	< 0.0001
YouAreWelcome	-0.458	-1.731	0.089
BCE2 : Face4	0.928	1.960	0.0547

Table 40: Significant associations with the likeability scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	4.117	28.911	< 0.0001
YouAreWelcome	-0.645	-3.325	0.002

Table 41: Significant associations with the likeability scores of robot characters by Ar participants.

c.2.6 *Intelligence of robot characters*

Term	Estimate	t value	Pr(> t)
Intercept	3.856	34.902	< 0.0001
YouAreWelcome	-0.349	-2.708	0.008

Table 42: Significant associations with the intelligence scores of robot characters by all participants.

Term	Estimate	t value	Pr(> t)
Intercept	3.646	13.154	< 0.0001
BCE2 : Face3	-0.884	-1.820	0.075
Face3 : male	-1.167	-1.866	0.064

Table 43: Significant associations with the intelligence scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	3.875	27.780	< 0.0001
YouAreWelcome	-0.439	-2.311	0.025

Table 44: Significant associations with the intelligence scores of robot characters by Ar participants.

C.2.7 *Safety of robot characters*

Term	Estimate	t value	Pr(> t)
Intercept	4.782	15.331	< 0.0001
AmE	-0.850	-2.124	0.036
male	-1.454	-6.128	< 0.0001
AmE : male	1.189	4.481	< 0.0001
AmE : BCE2 : Face4	1.282	1.707	0.09

Table 45: Significant associations with the safety scores of robot characters by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	3.833	16.968	< 0.0001
BCE2 : Face4	1.009	2.282	0.026

Table 46: Significant associations with the safety scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	4.694	30.495	< 0.0001
male	-1.254	-5.977	< 0.0001

Table 47: Significant associations with the safety scores of robot characters by Ar participants.

c.2.8 Safety of robot characters (calibrated)

These are the results for the safety scores with the pre-study questionnaire safety scores subtracted.

Term	Estimate	t value	Pr(> t)
Intercept	1.111	2.545	0.013
BCE2 : Face 2	-1.555	-1.781	0.078
AmE : BCE2 : Face 2	1.956	1.771	0.080
AmE : BCE2 : Face 4	2.022	1.763	0.081
BCE2 : Face 4 : male	2.306	1.928	0.057
AmE : BCE2 : Face 4 : male	-3.344	-2.093	0.039

Table 48: Significant associations with the calibrated safety scores of robot characters by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
BCE2 : Face4	0.906	1.922	0.059

Table 49: Significant associations with the calibrated safety scores of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	0.8203	3.971	< 0.0003
male	-0.918	-4.337	< 0.0001
YouAreWelcome	0.4779	2.258	0.029

Table 50: Significant associations with the calibrated safety scores of robot characters by Ar participants.

C.2.9 Locating the destination on the map

Backward stepwise regression results in a model with no significant terms. Fitting a model with linear terms for location and welcome suggests a trend of location C resulting in a better success rate: estimated coefficient is 1.350, $z = 1.856$, $p = 0.063$.

C.2.10 Thanking

Term	Estimate	t value	Pr(> t)
Intercept	2.042	16.180	< 0.0001
AmE	0.651	4.057	< 0.0001
male	0.923	5.365	< 0.0001
AmE : male	-0.901	-3.912	< 0.0002

Table 51: Significant associations with thanking of robot characters by all participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	2.800	16.382	< 0.0001
male	-0.514	-1.931	0.058
Face2 : male	1.070	2.808	0.007

Table 52: Significant associations with thanking of robot characters by AmE participants. Interactions are denoted by colons.

Term	Estimate	t value	Pr(> t)
Intercept	2.042	14.460	< 0.0001
male	0.923	4.795	< 0.0001

Table 53: Significant associations with thanking of robot characters by Ar participants.

BIBLIOGRAPHY

- Acapela Group. Acapela text-to-speech software. <http://www.acapela-group.com/>. Accessed December 12, 2012. (Cited on pages 83 and 113.)
- Michael Agar. *Language shock: Understanding the culture of conversation*. William Morrow, New York, 1994. (Cited on pages 5 and 12.)
- Michael Agar. Culture: Can you take it anywhere? Invited lecture presented at the Gevirtz Graduate School of Education, University of California at Santa Barbara. *Int. J. of Qualitative Methods*, 5(2), 2006. (Cited on pages 5 and 11.)
- Maria Jose Coperias Aguilar. Intercultural (mis)communication: The influence of L1 and C1 on L2 and C2. A tentative approach to textbooks. *Cuadernos de Filología Inglesa*, 7(1):99–113, 1998. (Cited on page 6.)
- Ainebot. Ainebot project home page. <http://distro.ibiblio.org/amigolinux/download/ainebot/>. Accessed: January 13, 2013. (Cited on page 111.)
- G. Allport. *The nature of prejudice*. Addison-Wesley, Cambridge, MA, 1954. (Cited on page 128.)
- Amna Alzeyara, Majd Sakr, Imran Fanaswala, and Nawal Behih. Realistic face and lip expressions for a bilingual humanoid robot. In *Proc. of Qatar Foundation Annual Research Forum*, page CSOS2, 2011. (Cited on page 113.)
- Jose Apesteguia, Ghazala Azmat, and Nagore Iriberri. The impact of gender composition on team performance and decision-making: Evidence from the field. *Management Science*, 58(1):78–93, 2012. (Cited on page 16.)
- Michael Argyle. *The psychology of interpersonal behavior*. Penguin, London, 1967. (Cited on page 25.)
- A. Atawneh. *Politeness theory and directive speech act in Arabic-English bilinguals: An empirical study (PhD thesis)*. State University of New York, Stony Brook, New York, 1991. (Cited on page 63.)
- Oliver Aubert and Yannick Prié. Advène: active reading through hypervideo. In *Proc. of ACM Hypertext*, September 2005. (Cited on page 31.)

- Nagham Awadallah. Palestinian and american phone conversation openings: Potential for intercultural miscommunication. In *Proceedings of 23rd Annual Symposium On Arabic Linguistics*, University of Milwaukee, Wisconsin, 2009. (Cited on page 54.)
- Benjamin Bailey. Communication of respect in interethnic service encounters. *Language in Society*, 26:327–356, 1997. (Cited on page 26.)
- Kathleen Bardovi-Harlig, Marda Rose, and Edelmira L. Nickels. The use of conventional expressions of thanking, apologizing, and refusing. In *Proceedings of the 2007 Second Language Research Forum*, pages 113–130, 2007. (Cited on pages 27, 54, and 62.)
- Ester Barinaga. ‘Cultural diversity’ at work: ‘National culture’ as a discourse organizing an international project group. *Human relations*, 60(2):315–340, 2007. (Cited on pages 17, 18, and 128.)
- Frederik Barth. *Ethnic groups and boundaries. The social organization of culture difference*. Universitetsforlaget, Oslo, 1969. (Cited on page 4.)
- Christoph Bartneck, Elizabeth Croft, and Dana Kulic. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In *Metrics for HRI Workshop, TR 471*, pages 37–44. Univ. of Hertfordshire, 2008. (Cited on page 95.)
- Christoph Bartneck, Elizabeth Croft, and Dana Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009. (Cited on pages 77, 86, 89, 101, and 124.)
- Joseph Bates. The role of emotion in believable agents. *Commun. ACM*, 37:122–125, July 1994. ISSN 0001-0782. doi: <http://doi.acm.org/10.1145/176789.176803>. URL <http://doi.acm.org/10.1145/176789.176803>. (Cited on page 8.)
- A. L. Baylor and Y. Kim. The role of gender and ethnicity in pedagogical agent perception. In *Proc. of E-Learn (World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education)*, Phoenix, Arizona, 2003. (Cited on page 18.)
- Amy L. Baylor and Yanghee Kim. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In James C. Lester, Rosa Maria Vicari, and Fábio Paraguacu, editors, *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, pages 268–270. Springer Berlin / Heidelberg, 2004. (Cited on pages 18 and 74.)
- Tilman Becker, Nate Blaylock, Ciprian Gerstenberger, Ivana Kruijff-korbayová, Andreas Korthauer, Manfred Pinkal, Michael Pitz, Peter

- Poller, and Jan Schehl. Natural and intuitive multimodal dialogue for in-car applications: The sammie system. In *In Proceedings of the ECAI Sub-Conference on Prestigious Applications of Intelligent Systems (PAIS 2006)*, Riva del Garda, 2006. (Cited on page 114.)
- Tara S. Behrend and Lori Foster Thompson. Similarity effects in on-line training: Effects with computerized trainer agents. *Computers in Human Behavior*, 27:1201–1206, 2011. (Cited on page 18.)
- Elizabeth S. Bell. *Theories of Performance*. Sage, Los Angeles, CA, 2008. (Cited on page 8.)
- P. M. Blau. *Inequality and heterogeneity: A primitive theory of social structure*. The Free Press, New York, 1977. (Cited on page 18.)
- Nate Blaylock and James Allen. A collaborative problem-solving model of dialogue. In *In Proceedings of the SIGdial Workshop on Discourse and Dialog*, pages 200–211, 2005. (Cited on page 114.)
- Marilynn B. Brewer. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86: 307–324, 1979. (Cited on page 15.)
- P. Brown and S. C. Levinson. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge, 1987. (Cited on pages 53 and 63.)
- Trung H. Bui. Multimodal dialogue management—state of the art. Technical report, Human Media Interaction Department, University of Twente, The Netherlands, January 2006. (Cited on page 114.)
- Chris Callison-Burch and Mark Dredze. Creating speech and language data with amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, CSLDAMT ’10, pages 1–12, Morristown, NJ, USA, 2010. Association for Computational Linguistics. URL <http://portal.acm.org/citation.cfm?id=1866696.1866697>. (Cited on page 60.)
- Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. Non-verbal cues for discourse structure. In *Proceedings of 41st Annual Meeting of ACL*, pages 106–115, Toulouse, France, 2001. (Cited on page 127.)
- Jr. Charles E. Johnson. Consistency of reporting of ethnic origin in the current population survey. Technical paper no. 31, Bureau of the Census, Washington, DC, 1974. (Cited on page 4.)
- Sonia Chernova, Jeff Orkin, and Cynthia Breazeal. Crowdsourcing HRI through online multiplayer games. In *Proc. of AAAI Fall Symposium on Dialog with Robots*, Washington, DC, 2010. AAAI. (Cited on page 60.)

- P. Collett. Training englishmen in the non-verbal behaviour of arabs. *International Journal of Psychology*, 6(3):209–215, 1971. (Cited on page 128.)
- Mark Cook. Gaze and mutual gaze in social encounters: How long—and when—we look others “in the eye” is one of the main signals of nonverbal communication. *American Scientist*, 65(2):328–333, 1977. (Cited on page 31.)
- P. T. Costa and R. R. McCrae. *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, FL, 1992. (Cited on page 66.)
- Helmut Daller and Cemal Yildiz. Power distance a work: The cases of Turkey, successor states of the former Soviet Union and Western Europe. *Journal of Politeness Research*, 2:35–53, 2006. (Cited on page 53.)
- Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proc. of Language Resources and Evaluation (LREC)*, 2006. (Cited on page 111.)
- Fiorella de Rosis, Catherine Pelachaud, and Isabella Poggi. Transcultural believability in embodied agents: A matter of consistent adaptation. In Sabine Payr and Robert Trapp, editors, *Agent Culture. Human-Agent Interaction in a Multicultural World*, pages 75–105. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey. London., 2004. (Cited on page 9.)
- Frédéric Delaunay and Joachim De Greeff. CONCEPT project. <http://www.tech.plym.ac.uk/SoCCE/CONCEPT/>, 2012. Accessed December 17, 2012. (Cited on page 112.)
- Anna-Marie Dew and Colleen Ward. The effects of ethnicity and culturally congruent and incongruent nonverbal behaviors on interpersonal attraction. *J. of Applied Social Psychology*, 23(17):1376–1389, 1993. (Cited on pages 18, 109, and 128.)
- Nancy Dorian. *Language Depth*. University of Pennsylvania Press, Philadelphia, 1981. (Cited on page 28.)
- Paul Ekman and W. Friesen. *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978. (Cited on pages xvi, 55, 112, and 120.)
- Steve L. Ellyson, John F. Dovidio, Randi L. Corson, and Debbie L. Vinicur. Visual dominance behavior in female dyads: Situation and personality factors. *Social Psychology Quarterly*, 43:328–335, 1980. (Cited on page 25.)

- R. J. Ely and D. A. Thomas. Cultural diversity at work: The effects of diversity perspectives on work group processes and outcomes. *Administrative Science Quarterly*, 46:229–273, 2001. (Cited on page 18.)
- R. V. Exline, S. L. Ellyson, and B. Long. Visual behavior as an aspect of power role relationships. In P. Pliner, L. Kramers, and T. Alloway, editors, *Advances in the study of communication and affect*, pages 21–52. Plenum, New York, 1975. (Cited on page 24.)
- Said Hassan Farahat. *Politeness phenomena in Palestinian Arabic and Australian English: A cross-cultural study of selected contemporary plays (PhD thesis)*. Australian Catholic University, Australia, 2009. (Cited on page 63.)
- Ellen Feghali. Arab cultural communication patterns. *International Journal of Intercultural Relations*, 21(3):345–378, 1997. (Cited on pages 7, 26, 54, and 62.)
- T. B. Fitzpatrick. Soleil et peau. *J. de Medecine Esthetique*, 2:33–34, 1975. (Cited on page 75.)
- Mohammed Ghawi. Pragmatic transfer in Arabic learners of English. *El Two Talk*, 1(1):39–52, 1993. (Cited on pages 27, 54, 55, 56, and 62.)
- Joseph C. Giarratano and Gary D. Riley. *Expert Systems: Principles and Programming*. Course Technology, 1998. (Cited on page 115.)
- A. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, 2002. (Cited on page 69.)
- Samuel D. Gosling, Peter J. Rentfrow, and Jr. William B. Swann. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37:504–528, 2003. (Cited on pages 27, 61, and 66.)
- Arthur Graesser, Vasile Rus, Sidney D’Mello, and G. Tanner Jackson. *AUTOTUTOR. Learning through natural language dialogue that adapts to the cognitive and affective states of the learner*. Information Age Publishing, 2008. (Cited on page 1.)
- B. J. Grosz and C. L. Sidner. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in communication*, pages 417–444. MIT press, 1990. (Cited on pages 113 and 114.)
- Terry Halfhill, Eric Sundstrom, Jessica Lahner, Wilma Calderone, and Tjai M. Nielsen. Group personality composition and group effectiveness. an integrative review of empirical research. *Small Group Research*, 36(1):83–105, February 2005. (Cited on page 17.)

- J. A. Hall, J. D. Carter, and T. G. Horgan. Gender differences in non-verbal communication of emotion. In A. Fischer, editor, *Gender and emotion: Social psychological perspectives*, pages 97–117. Cambridge University Press, Paris, 2000. (Cited on page 108.)
- Maureen T. Hallinan and Warren N. Kubitschek. Sex and race effect of the response to intransitive sentiment relations. *Social Psychology Quarterly*, 53(3):252–263, 1990. (Cited on page 16.)
- K. Hayashi, T. Kanda, T. Miyashita, H. Ishiguro, and N. Hagita. Robot manzai—robots’ conversation as a passive social medium. In *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pages 456–462, December 2005. (Cited on page 9.)
- B. Hayes-Roth, H. Maldonado, and M. Moraes. Designing for diversity: Multi-cultural characters for a multi-cultural world. In *Proc. of IMAGINA*, pages 207–225, Monte Carlo, Monaco, 2002. (Cited on pages 19 and 20.)
- Chiara Higgins, Elizabeth McGrath, and Lailla Moretto. Mturk crowdsourcing: a viable method for rapid discovery of arabic nicknames? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, CSLDAMT ’10, pages 89–92, Morristown, NJ, USA, 2010. Association for Computational Linguistics. URL <http://portal.acm.org/citation.cfm?id=1866696.1866710>. (Cited on pages 60 and 67.)
- Jerry R. Hobbs and Alicia Sagae. A commonsense theory of microsociology: Interpersonal relationships. In *Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning, AAAI Spring Symposium Series*, 2011. (Cited on page 63.)
- Geert Hofstede. *Culture’s Consequences: comparing values, behaviors, institutions, and organizations across nations*. SAGE Publications, Thousand Oaks, California, 2001. (Cited on pages 19, 24, and 39.)
- Janet Holmes and Maria Stubbe. *Power and Politeness in the Workplace*. Longman, London, 2003. (Cited on page 53.)
- Sander Hoogendorn, Hessel Oosterbeek, and Mirjam van Praag. The impact of gender diversity on the performance of business teams: evidence from a field experiment. Discussion paper, Tinbergen Institute, Amsterdam School of Economics, University of Amsterdam, April 2011. (Cited on page 16.)
- M. E. Hoque, L-P. Morency, and R. W. Picard. Are you friendly or just polite? —Analysis of smiles in spontaneous face-to-face interactions. In *Proc. of the Affective Computing and Intelligent Interaction*, October 2011. (Cited on pages 26 and 30.)

- Jeff Howe. Crowdsourcing: A definition. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html, 2006. Accessed January 20, 2013. (Cited on page 59.)
- Daniel B. Hrdy. Analysis of hair samples of mummies from semna south (sudanese nubia). *American Journal of Physical Anthropology*, 49(2):277–282, 1978. (Cited on page 75.)
- Francisco Iacobelli and Justine Cassell. Ethnic identity and engagement in embodied conversational agents. In *Proc. Int. Conf. on Intelligent Virtual Agents*, pages 57–63, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-74996-7. (Cited on pages 6, 9, 18, and 20.)
- John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005. (Cited on pages 108 and 109.)
- Amy Isard, Carsten Brockmann, and Jon Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the Fourth International Natural Language Generation Conference, INLG '06*, pages 25–32, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-72-8. (Cited on page 63.)
- Katherine Isbister and Clifford Nass. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. J. Hum.-Comput. Stud.*, 53:251–267, August 2000. ISSN 1071-5819. doi: 10.1006/ijhc.2000.0368. URL <http://portal.acm.org/citation.cfm?id=350073.350077>. (Cited on page 17.)
- Hiroshi Ishiguro. Android science. toward a new cross-interdisciplinary framework. In *Proc. of Cog. Sci. Workshop: Toward Social Mechanisms of Android Science*, 2005. (Cited on pages 8, 19, and 108.)
- K.R. Johnson. Black kinesics: Some non-verbal communication patterns in the black culture. In L. A. Samovar and R. E. Porter, editors, *Intercultural communication: A reader*, pages 259–568. Wadsworth Publishing Company, Belmont, California, 1976. (Cited on page 25.)
- Lewis W. Johnson and Andre Valente. Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In *AAAI*, pages 1632–1639, 2008. (Cited on page 127.)
- S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey. Anthropomorphic interactions with a software agent and a robot. *Social Cognition*, 26(2):168–180, 2008. (Cited on page 124.)
- Kyunghee Kim. *Affect Reflection Technology in Face-to-Face Service Encounters*. MIT MS Thesis, September 2009. (Cited on page 25.)

- S. Kita, I. van Gijn, and H. van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachs-muth and M. Fröhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 23–35. Springer, 1998. (Cited on page 30.)
- Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: <http://doi.acm.org/10.1145/1357054.1357127>. URL <http://doi.acm.org/10.1145/1357054.1357127>. (Cited on page 60.)
- David D. Laitin. What is a language community? *American Journal of Political Science*, 44(1):142–155, 2000. (Cited on pages 28 and 101.)
- P. Lazarsfeld and R. Merton. Friendship as a social process: a substantive and methodological analysis. In *Freedom and Control in modern society*. Van Nostrand, New York, 1954. (Cited on page 1.)
- Dale G. Leathers. *Successful nonverbal communication: principles and applications*. Allyn and Bacon, Boston, 1997. (Cited on page 7.)
- Eun-Ju Lee. Effects of “gender” of the computer on informational social influence: the moderating role of task type. *Int. J. Human-Computer Studies*, 58:347–362, 2003. (Cited on page 16.)
- Eun-Ju Lee, Clifford Nass, and Scott Brave. Can computer-generated speech have gender? an experimental test of gender stereotypes. In *Int. Conf. on Computer-Human Interaction (CHI)*, pages 289–290, New York, 2000. ACM Press. (Cited on pages 1 and 16.)
- Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *J. of Communication*, (56):754–772, 2006. (Cited on page 17.)
- Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *Proc. Int. Conf. on Human-Robot Interaction*, pages 203–210. ACM, 2010a. ISBN 978-1-4244-4893-7. (Cited on page 60.)
- Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Receptionist or information kiosk: how do people talk with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW*, pages 31–40, New York, NY, USA, 2010b. ACM. ISBN 978-1-60558-795-0. (Cited on page 102.)

- Richard M. Levinson, Kelly T. McCollum, and Nancy G. Kutner. Gender homophily in preferences for physicians. *Sex roles*, 10(5–6):315–325, 1984. (Cited on page 16.)
- Chyi-Yeu Lin, Chang-Kuo Tseng, Wei-Chung Teng, Wei-Chen Lee, Chung-Hsien Kuo, Hung-Yan Gu, Kuo-Liang Chung, and Chin-Shyurng Fahn. The realization of robot theater: Humanoid robots and theatric performance. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–6, June 2009. (Cited on page 9.)
- K. E. Lockbaum. A collaborative planning model of intentional structure. *Computational linguistics*, 24(4):525–572, 1998. (Cited on page 113.)
- D. A. Mackey, C. H. Wilkinson, L. S. Keams, and A. W. Hewitt. Classification of iris colour: review and refinement of a classification schema. *Royal Australian and New Zealand College of Ophthalmologists*, 39(5):462–471, 2011. (Cited on page 75.)
- Francois Mairesse and Marilyn Walker. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proc. of 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008. (Cited on pages 66, 122, and 123.)
- Francois Mairesse and Marilyn Walker. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278, 2010. (Cited on pages 63, 66, and 127.)
- Maxim Makatchev and Reid Simmons. Incorporating a user model to improve detection of unhelpful robot answers. In *Proc. of RO-MAN*, Toyama, Japan, September–October 2009. (Cited on page 102.)
- Maxim Makatchev and Reid Simmons. Interaction manager. <http://github.com/maxipesfix>, 2013. To be released. (Cited on pages 114 and 116.)
- Maxim Makatchev, Min Kyung Lee, and Reid Simmons. Relating initial turns of human-robot dialogues to discourse. In *Proc. of the Int. Conf. on Human-Robot Interaction (HRI)*, pages 321–322. ACM, 2009. (Cited on page 102.)
- Maxim Makatchev, Reid Simmons, and Majd Sakr. Carnegie Mellon Receptionist Corpus. <http://www.qatar.cmu.edu/hala/corpora/>, 2012. (Cited on pages 26 and 57.)
- Normal N. Markel, Judith A. Phillis, Robert Vargas, and Kenneth Hoard. Personality traits associated with voice types. *J. of Psycholinguistic Research*, 1(3):249–255, 1972. (Cited on page 74.)

- Marianne Schmid Mast and Judith A. Hall. Who is the boss and who is not? Accuracy in judging status. *Journal of Nonverbal Behavior*, 28: 145–165, 2004. (Cited on page 24.)
- N. Mavridis and D. Hanson. The IbnSina interactive theater: Where humans, robots and virtual characters meet. In *Proc. of RO-MAN*, Toyama, Japan, 2009. (Cited on page 19.)
- Evelyn McClave, Helen Kim, Rita Tamer, and Milo Mileff. Head movements in the context of speech in arabic, bulgarian, korean, and african-american vernacular english. *Gesture*, 7(3):343–390, 2007. (Cited on page 25.)
- M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. (Cited on pages 15 and 16.)
- Enid Montague, Jie Xu, Ping yu Chen, Onur Asan, Bruce P. Barret, and Betty Chewning. Modeling eye gaze patterns in clinician-patient interaction with lag sequential analysis. *J. of Human Factors and Ergonomics Society*, 53:502–516, October 2011. (Cited on page 25.)
- Biren (Ratnesh) A. Nagda. Breaking barriers, crossing borders, building bridges: communication processes in intergroup dialogues. *Journal of Social Issues*, 62(3):553–576, 2006. (Cited on page 128.)
- Joane Nagel. Constructing ethnicity: Creating and recreating ethnic identity and culture. *Social problems*, 41(1):152–176, 1994. (Cited on pages 4 and 11.)
- Sharon Nagi. Making room for migrants, making sense of difference: spatial and ideological expressions of social diversity in urban Qatar. *Urban Studies*, 43(1):119–137, 2006. (Cited on page 128.)
- C. Nass, K. Isbister, and E. Lee. Truth is beauty: Researching embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied conversational agents*, pages 374–402. MIT Press, Cambridge, MA, 2000. (Cited on pages 1 and 18.)
- Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. Evaluating the effect of gesture and language on personality perception in conversational agents. In *Proceedings of the 10th international conference on Intelligent virtual agents*, IVA'10, pages 222–235, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15891-9, 978-3-642-15891-9. URL <http://portal.acm.org/citation.cfm?id=1889075.1889103>. (Cited on page 123.)
- Gayle L. Nelson, Magmoud Al Batal, and Waguida El Bakary. Cross-cultural pragmatics: strategy use in Egyptian Arabic and American

- English refusals. *Applied Linguistics*, 23:163–189, 2002. (Cited on page 57.)
- Margaret Nydell. *Understanding Arabs: A guide for Westerners*. Intercultural press, 1987. (Cited on page 24.)
- Kohei Ogawa, Koichi Taura, and Hiroshi Ishiguro. Possibilities of androids as poetry-reciting agent. In *RO-MAN, 2012 IEEE*, pages 565–570, September 2012. (Cited on page 9.)
- Eunil Park, Dallae Jin, and Angel P. del Pobil. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9(35), 2012. (Cited on pages 1 and 17.)
- Catherine Pelachaud. Multimodal expressive embodied conversational agents. In *Proc. of ACM Multimedia*, pages 683–689, 2005. (Cited on page 21.)
- Thomas F. Pettigrew. Intragroup contact theory. *Annu. Rev. Psychol.*, 49:65–85, 1998. (Cited on page 128.)
- Thies Pfeiffer. Understanding multimodal deixis with gaze and gesture in conversational interfaces. Doctoral dissertation, Bielefeld University, 2010. (Cited on page 44.)
- Katherine W. Phillips and Damon J. Phillips. Heterogeneity, performance, and blau s paradox: The case of nhl hockey teams, 1988–1998. *Manuscript in development*. (Cited on page 18.)
- J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000. (Cited on pages 67 and 139.)
- Yaniv Poria. Gender—a crucial neglected element in the service encounter: An exploratory study of the choice of hotel masseur or masseuse. *Journal of Hospitality and Tourism Research*, 32(2):151–168, 2008. (Cited on page 16.)
- B. Reeves and Cliff Nass. *The Media Equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996. (Cited on page 1.)
- Matthias Rehm, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner, Yukiko Nakano, Afia Akhter Lipi, Toyooki Nishida, and Hung-Hsuan Huang. Creating standardized video recordings of multimodal interactions across cultures. In Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen, editors, *Multimodal corpora*, pages 138–159. Springer-Verlag, Berlin, Heidelberg, 2009a. (Cited on pages 21 and 26.)
- Matthias Rehm, Yukiko Nakano, Elisabeth André, Toyooki Nishida, Nikolaus Bee, Birgit Endrass, Michael Wissner, Afia Akhter Lipi,

- and Hung-Hsuan Huang. From observation to simulation: generating culture-specific behavior for interactive systems. *AI and Society*, 24(3):267–280, 2009b. (Cited on page 21.)
- Reid Simmons. Interprocess communication (ipc). <http://www.cs.cmu.edu/~IPC>, 2012. Accessed: January 13, 2013. (Cited on page 111.)
- T. Ribeiro and A. Paiva. The illusion of robotic life: Principles and practices of animation for robots. In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pages 383–390, March 2012. (Cited on page 8.)
- C. Rich, A. Holroyd, B. Ponsler, and C. Sidner. Recognizing engagement in human-robot interaction. In *Proceedings of ACM/IEEE International Conference on Human Robot Interaction*, pages 375–382, 2010. (Cited on page 49.)
- Charles Rich and Candace L. Sidner. Collagen: A collaboration manager for software interface agents. *Journal of User Modeling and User-Adapted Interaction*, 8:315–350, 1998. (Cited on page 114.)
- Charles Rich and Candace L. Sidner. Using collaborative discourse theory to partially automate dialogue tree authoring. In *Proc. Int. Conf. on Intelligent Virtual Agents*, 2012. (Cited on page 114.)
- Cecilia L. Ridgeway, Joseph Berger, and LeRoy Smith. Nonverbal cues and status. *American Journal of Sociology*, 90(5):955–978, 1985. (Cited on page 24.)
- Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-930-5. doi: <http://doi.acm.org/10.1145/1753846.1753873>. URL <http://doi.acm.org/10.1145/1753846.1753873>. (Cited on pages 60 and 67.)
- J. Philippe Rushton. Ethnic differences in temperament. In Yueh-Ting Lee, Clark R. McCauley, and Juris G. Draguns, editors, *Personality and person perception across cultures*, pages 45–63. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey, 1999. (Cited on page 63.)
- Patrick B. Ryan, David Madigan, Paul E. Stang, J. Marc Overhage, Judith A. Racoosin, and Abraham G. Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in Medicine*, 31(30):4401–4415, 2012. ISSN 1097-0258. (Cited on page 109.)

- Eman Saadah. The how are you? sequence in telephone openings in Arabic. *Proceedings of Illinois Language and Linguistics Society 1: Language online*, pages 171–186, 2009. (Cited on page 54.)
- M. Safadi and C. A. Valentine. Contrastive analyses of american and arab nonverbal and paralinguistic communication. *Semiotica*, 83(3/4):269–292, 1990. (Cited on page 25.)
- Klaus R. Scherer. Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality*, 40:191–210, 1972. (Cited on pages 69 and 74.)
- Klaus P. Schneider. Genre matters: Textual and contextual constraints on contemporary English speech behaviour. *Anglia — Zeitschrift fur englische Philologie*, 125(1):59–83, 2007. (Cited on page 6.)
- Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel van Aken, and Win Meeus. Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of Personality*, 78(2):509–538, 2010. (Cited on pages 15 and 17.)
- Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. Persuasive robotics: The influence of robot gender on human behavior. In *Int. Conf. on Intelligent Robots and Systems*, pages 2563–2568, St. Louis, USA, 2009. IEEE/RSJ. (Cited on page 16.)
- R. Simmons, M. Makatchev, R. Kirby, M.K. Lee, I. Fanaswala, B. Browning, J. Forlizzi, and M. Sakr. Believable robot characters. *AI Magazine*, 32(4):39–52, 2011. (Cited on pages 2 and 12.)
- Konstantin Stanislavski. *An Actor's Work: A Student's Diary*. Routledge, New York, NY, 2008. (Cited on pages 122 and 123.)
- Maria Staudte and Matthew W. Crocker. The effect of robot gaze on processing robot utterances. In *Proc. 31st Annual Conf. of the Cognitive Science Society*, Amsterdam, 2009. (Cited on page 1.)
- K. T. Strongman and B. G. Champness. Dominance hierarchies and conflict in eye contact. *Acta Psychologica*, 28:376–386, 1968. (Cited on page 25.)
- Adriana Tapus, Cristian Tapus, and Maja J. Mataric. User-robot personality matching and robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics Journal*, 1(2):169–183, 2008. (Cited on page 17.)
- F. Thomas and O. Johnston. *Disney Animation: The illusion of life*. Abbeville Press, New York, 1981. (Cited on page 8.)
- Mark Conway Thompson. Personal communication, 2012. (Cited on page 44.)

- Richard Todd. Speaker-ethnicity: Attributions based on the use of prosodic cues. In *Proc. 1st Int. Conference on Speech Prosody*, 2002. (Cited on page 74.)
- Véronique Traverso. Syrian service encounters: a case of shifting strategies within verbal exchange. *Pragmatics*, 11(4):421–444, 2001. (Cited on page 26.)
- Nancy Brandom Tuma and Maureen T. Hallinan. The effects of sex, race, and achievement on schoolchildren's friendships. *Social Forces*, 57(4):1265–1285, 1979. (Cited on page 16.)
- A. J. Vaccaro. Personality clash. *Personnel Administrator*, 33:88–92, 1988. (Cited on page 17.)
- B. Velichkovsky, A. Sprenger, and M. Pomplun. Auf dem weg zur blickmaus: Die beeinflussung der fixationsdauer durch kognitive und kommunikative aufgaben. *Software-Ergonomie*, 1997. (Cited on page 44.)
- Luis von Ahn and Laura Dabbish. General techniques for designing games with a purpose. *Communications of the ACM*, pages 58–67, August 2008. (Cited on pages 61 and 127.)
- Mario von Cranach and Johann H. Ellgring. Problems in the recognition of gaze direction. In M. von Cranach and I. Vine, editors, *Social Communication and Movement: Studies of Interaction and Expression in Man and Chimpanzee*, pages 419–443. Academic Press, London, 1973. (Cited on page 31.)
- Nigel G. Ward and Yaffa Al Bayyari. A prosodic feature that invites back-channels in egyptian arabic. In Mustafa Mughazy, editor, *Perspectives on Arabic Linguistics*, pages 187–206. John Benjamins, 2007. (Cited on page 126.)
- Larry Wasserman. Most findings are false. <http://normaldeviate.wordpress.com/2012/12/27/most-findings-are-false/>, 2012. Accessed January 16, 2013. (Cited on page 109.)
- Mary C. Waters. *Ethnic Options: Choosing Identities in America*. University of California Press, 1990. (Cited on page 4.)
- D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. of Personality and Social Psychology*, 47:1063–1070, 1988. (Cited on page 27.)
- W. E. Watson, K. Kumar, and L.K. Michaelson. Cultural diversity s impact on interaction process and performance: Comparing homogeneous and diverse task groups. *Academy of Management Journal*, 36:590–602, 1993. (Cited on page 17.)

- Suzanne Wertheim. Personal communication, 2013. (Cited on page 53.)
- K. Y. Williams and C. A. O'Reilly. Demography and diversity in organizations: A review of 40 years of research. *Research in organizational behavior*, 20:77–140, 1998. (Cited on page 17.)
- Sarah Woods, Kerstin Dautenhahn, Christina Kaouri, René te Boekhorst, and Kheng Lee Koay. Is this robot like me? links between human and robot personality traits. In *Int. Conf. on Humanoid Robots (Humanoids)*, pages 375–380, Tsukuba, Japan, 2005. IEEE-RAS. (Cited on page 17.)
- Langxuan Yin, Timothy Bickmore, and Donna Byron. Cultural and linguistic adaptation of relational agents for health counseling. In *Proc. of CHI*, Atlanta, Georgia, USA, 2010. (Cited on page 21.)