A Computational Approach to Etiquette and Politeness:
Validation Experiments

Christopher A. Miller, Peggy Wu, Harry B. Funk

Smart Information Flow Technologies
211 First St. N. Ste. 300
Minneapolis, MN 55401
(612) 339-7438, (612) 339-7437
{cmiller, pwu, hfunk}@sift.info

Abstract

A significant and prevalent aspect of human behavior (as opposed to attitudes and beliefs) which is affected by cultural differences is communication style. We have been developing computational models of an important culturally-varying aspect of communication: "politeness" and "etiquette" in social interactions and its role in establishing and managing power and familiarity relationships, urgency, indebtedness, etc. A valid computational model of such interactions would enable the creation of better simulations and games for language and culture training, as well as aid in the design of materials and machines to better serve members of a given culture. We have developed such a model based on a rich, universal theory of human-human politeness and "etiquette". This model links observable and inferred aspects of social context (power and familiarity relationships, imposition, knowledge about character), which have culture-specific values, to produce expectations about politeness behaviors (also culturally defined). By using observations of politeness behaviors (or their lack), the same model permits inferences and updates about those attributes. We briefly describe the algorithm we have developed and describe results from two validation experiments involving first trained, and later untrained raters. We have used this model in other work to guide simulated game agents in interpreting etiquette directed at them and in generating politeness behaviors in response. While other methods of interactive behavior generation are available (e.g., behavior scripting) our modular, computational approach shows promise for reducing software development costs and/or increasing the breadth of an agent's social interaction behavior through the creation of modular, cross-cultural etiquette libraries.

Introduction

Our interest in cultural differences and similarities is primarily focused on interactions between members of different cultures. While cultural differences and similarities may be of interest in their own right, that interest is somewhat academic. It becomes practical when we need to get a job done with team members coming from different cultures, or we need to communicate information and obtain resulting actions from members of another culture. This makes apparent the role of *social interaction*—that is, interactions based on the social characteristics and assumptions of each agent as an intentional entity (Dennett 1989) and drawing from culturally familiar patterns of expectations about appropriate behaviors among such agents in cross cultural studies. While the study of culturally-correlated attitudes, cognitive styles, and sense-making mechanisms, not to mention specific attributes and histories of alternate cultures, are all important, action in a multi-cultural exchange almost inevitably involves communication—and that frequently (though not always) means verbal and non-verbal exchange between intelligent agents by means of language and gesture.

There has been much theoretical and basic research on identifying cultural patterns (Hofstede, 2001) and how cultural factors affect cognitive processes (Nisbett, 2003), but none provides a direct link of these cultural factors to human performance, nor is readily amenable to computational modeling at a fine level of granularity. The research to date has provided valuable insights on how to structure and 'quantify' culture, but has rarely resulted in concrete, quantitative models which predict or account for human behavior—and especially not at the micro-level of human communication. Cultural factors, mingled with situational context and personal preferences, manifest into human behaviors that are difficult to analyze and tease apart.

An additional challenge lies in finding performance phenomenon that are both predictable from cultural factors and are worth predicting—that is, that have valuable outcomes. There is little doubt that cultural factors do affect performance. For example, Nisbett has found that North Americans and South East Asians see different objects in the same picture due to what he calls field dependence (Nisbett, 2003), implying cognitive differences in pattern recognition, problem solving, and decision making skills among cultures, all of which contribute to performance. But it has proven difficult to trace the chain of causality

from these differences to actual, valuable behavioral differences.

We have instead focused on developing a concrete model of observable human behaviors—specifically communication behaviors having to do with etiquette and politeness—which in turn have relevance for human performance, attitudes and broader decision making.

"Politeness" for Social Interactions?

The terms "etiquette" and "politeness" are likely to evoke notions of dinner forks and curtsies. But *politeness* is a well-studied anthropological, sociological and linguistic phenomenon. It is the processes by which we determine and manage the "threat" inherent in communication between intentional actors which are presumed to have goals and the potential to take offense at having those goals thwarted (cf. Dennet, 1989; Goffman, 1967). Politeness is one means by which we signal, interpret, maintain and alter social relationships. *Etiquette* is the code by which we signal politeness. It makes use of verbal, physical, gestural and even more primitive modes of interaction. For example, deference can be expressed by posture, by quiet speech and/or by explicit markers such as titles and honorifics. The key is the set of cultural expectations which allow interactants to interpret the behavior, or lack of behavior, in a predetermined fashion. In this sense, there is a "cultural etiquette" associated with, say, infantry soldiers as opposed to clerical workers, just as there is a one for marketplace negotiations in the Middle East vs. an American shopping center.

As such, therefore, politeness and etiquette (at least in our sense) should be very much at the forefront of training and managing social interactions, and they should play a large role in determining the believability and effectiveness of simulations and training materials. Believable behavior is behavior that is understandable (i.e., the viewer can infer intent behind the behavior) and broadly consistent with a domain expert's (e.g., native speaker's) expectations. Understandability and expectations, in turn, depend upon the social and cultural context of the behavior. Etiquette provides a way of modeling interactions and moves within a social and cultural context, and of predicting their impact on observers' interpretations about the motives, understanding, knowledge and relationships of those who exhibit them. Therefore, we have focused on modeling etiquette and its role in achieving believability. If simulated characters do not behave in accordance with etiquette-based expectations, one of two outcomes may result: either (1) they will not be perceived as believable, or (2) they will be misinterpreted—the trainee will draw false inferences about their relationships, intentions, etc. In either event, they will be useless for training purposes—and worse yet, they may produce inaccurate expectations in students who interact with them.

A Model of Human-Human Etiquette for Politeness

A seminal body of work in the study of politeness is the cross-cultural studies and resulting model developed by Brown and Levinson (1987). Brown and Levinson noted that people across cultures and languages regularly depart from strictly efficient conversation by using conversational behaviors designed to mitigate or soften direct expressions of desire, intent or command. A simple illustation in English would be: as we settle down to a meal together, I ask you "Please pass the salt." The use of "please" in that sentence is unnecessary for a truthful, relevant or clear expression of my wish and is, in fact, not required to express my overt intent. Over years of cross linguistic and cross cultural studies, Brown and Levinson collected a huge database of such violations, and developed a model to explain their occurrence, which will be explained next.

Face threats in social interactions

The Brown and Levinson model assumes that social actors are motivated by two important social wants based on the concept of face (Goffman, 1967) or, loosely, the "positive social value a person effectively claims for himself" (cf. Cassell and Bickmore, 2002, p. 6). Face can be "saved" or lost, and it can be threatened or conserved in interactions. Brown and Levinson further refine the concept of face into two specific subgoals that all social actors can be presumed to have:

1. *Positive face*—an individual's desire to be held in high esteem, to have his/her actions and opinions valued, to be approved of by others, etc.
2. *Negative face*—an individual's desire for autonomy, to have his/her will, to direct his/her attention where and when desired, etc.

Virtually all interactions between social agents are to some degree Face Threatening Acts (FTAs). My simple act of speaking to you, regardless of the content of my words, places a demand on your attention that threatens your negative face, for example. This, then, is the reason for the "please" in my request for salt: If I simply state my desire as bald propositional content (e.g., "Give me the salt") I would be ambiguous about whether I have the power or right to compel you to give me salt. You might well take offense at the implication. The "please" is thus a "redressive" strategy which mitigates the threat. Furthermore, the expectation that such a strategy be used is an example of etiquette that enables interpretations. The etiquette is the "rule" that entitles us to conclude that those who use "please" are striving to play by the rules—striving to be seen as polite; those who do not are not striving to be polite for various reasons (perhaps they don't believe they need to be, perhaps their notions about politeness are different, perhaps they are just rude).

Computing the severity of a face threat

The core of Brown and Levinson's model is the claim that the degree of face threat posed by an act is provided by the function:

$$W_x = P(H,S) + D(S,H) + R_x$$

- Where, W_x is the 'weightiness' or severity of the FTA x
- $P(H,S)$ is the relative power that H has over S. Power is an asymmetric relationship.
- $D(S,H)$ is the social distance between the speaker (S) and the hearer (H). Social distance is roughly the inverse of familiarity and is a symmetric relationship
- R_x is the ranked imposition of the raw act itself.

Brown and Levinson themselves do not operationalize these parameters; instead, they are offered as qualitative constructs. Work by Cassell and Bickmore (2003) and by Johnson and Rizzo (2004) has created numerical representations for them toguide, respectively, a simulated real estate agent in making small talk and a pedagogical agent in offering advice and criticism. Our goal has been to develop a computational formulation of the Brown and Levinson algorithm for use in free-flowing conversation and social interactions between humans and agents in a simulation environment.

In our implementation of Brown and Levinson, described below, we add an additional term. They allude to, but don't explicitly include, a factor representing the relative weighting an individual puts on his/her own goals vs. the face goals of others-- his/her general willingness to place others' needs first. We have called that term "character" and introduce a term for it, abbreviated as C(S) for the character of speaker (S).

Redressing Face Threats

Since FTAs are potentially disruptive to human-human relationships, we generally make use of redressive strategies to mitigate the degree of face threat imposed by our actions. The core of Brown and Levinson's model is the claim that the degree of face threat posed by an act must be redressed or balanced by the value of the politeness behaviors used if the social status quo is to be maintained. That is:

- $W_x \cong V(\mathbf{A}_x)$
- Where W_x is the "weightiness" of severity of a face threat x, and
- $V(\mathbf{A}_x)$ is the combined redressive value of the set of politness behaviors ($\mathbf{A}_x$) used in the interaction.

If less redress is used than is perceived as necessary, that is if $W_x \gg V(\mathbf{A}_x)$, then the utterance will be perceived as rude and the hearer may seek alternative explanations or interpretations for the behaviors, as will be discussed below. If more politeness behaviors are used than are perceived as necessary, that is if $W_x \ll V(\mathbf{A}_x)$, then the utter-
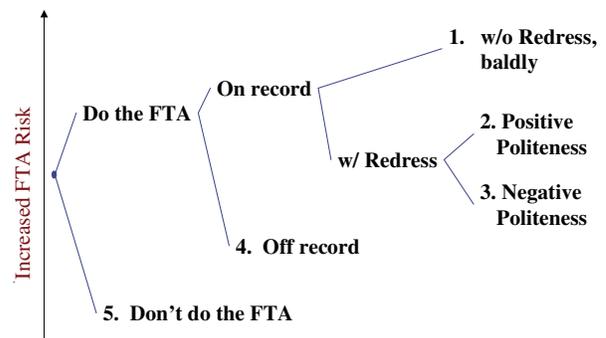

Figure 1. Universal redress strategies as ranked by Brown and Levinson (1987).

ance will be perceived as "over-polite" or obsequious and, again, ulterior motives for the behaviors or ulterior interpretations of the context may be sought.

Brown and Levinson offer an extensive catalogue of universal strategies for redressing, organized according to 5 broad strategies. These are illustrated in Figure 1 ranked from least to most threatening.

- The least threatening approach is simply not to do the FTA. Some FTAs simply can't be performed without insult, regardless of the amount of redress offered.
- Off record FTA strategies are means of doing the act with "plausible deniability" by means of innuendo and hints. An "off record" method of asking for salt might be "I find this food a bit bland."
- Overt FTA can still be mitigated by offering redress aimed at either positive or negative face. Brown and Levinson suggest that negative redress will be more effective. Negative redressive strategies focus on H's negative face needs—independence of action and attention. They minimize the impact on H by being direct and simple in making the request, offering apologies and deference, minimizing the magnitude of the imposition and/or explicitly incurring a debt.
- Positive redressive strategies target the hearer's positive face. These strategies emphasize common ground between S and H by noticing and attending to H, by invoking in-group identity, by joking and assuming agreement and/or by explicitly offering rewards/promises.
- Finally, the most threatening way to perform an FTA is "baldly, on record," without any redress.

Brown and Levinson's model doesn't stop at that level, however. For positive and negative redressive and off record strategies, they offer a host of well-researched examples from at least three different language/culture groups (English, Tamil and Tzeltal) organized into a structure of mutually supporting and incompatible approaches. We do not have space to present their findings in depth, but we note as an example that their categorization of negative redress strategies contains 10 alternate approaches, some of which are mutually supporting or conflicting, including:

- Be Pessimistic—"You're not going to pass me the salt, are you?"
- Minimize the Imposition—"Could you just nudge that salt shaker over here?"
- Give Deference—"Excuse me, sir, would you pass the salt?"
- Apologize—"I'm sorry to interrupt, but would you pass the salt?"

An "Etiquette Metric"—Believable levels of Politeness

According to the Brown and Levinson model described above, people generally want to accomplish their goals expeditiously-- and this argues for minimizing redressive strategies. But they also experience a range of social and personal pressures to not threaten the face of those they interact with (especially those with greater power or shared familiarity)-- and this argues for extensive redressive strategies. The balance between these pressures yields the selection of specific strategies in context.

As stated above, the core of the Brown and Levinson model is the notion that the relationship between face threat weight and amount or value of redressive behaviors used determines whether an interaction is perceived as nominal, rude or overly polite. Of course, an individual's perception of the rudeness of an interaction will, in turn, be dependent on that individual's perceptions of the face threat and redress involved. We have expressed this relationship as follows. Perceived politeness is a function of the Perceived Imbalance (I) between the perceived degree of face threat in an interaction and the perceived amount or degree of redress used. We can express this as:

$$B_O:I_x = B_O:V(\mathbf{A}_x) - B_O:W_x$$

Where B_O is the Belief of Observer O about the other terms in the equation and I_x is the perceived Imbalance (I) of interaction x. Thus, this equation says that the believed imbalance as perceived by Observer O of interaction x will be the difference between the value of the redressive acts A in x (as perceived by O) minus the amount of face threat W (as perceived by O)—which is itself a function of the Observer's beliefs about the P, D, and R terms defined above. Imbalance will be positive when more redressive politeness behaviors were used than there was face threat present—corresponding to the overly polite or obsequious condition. I_x will be negative when less redress is used than there was threat—a rude condition.

This model also explains a fundamental issue about politeness use—namely, the fact that the same set of politeness behaviors, used in different contexts, may well be perceived as anything from appropriate to rude or over-polite. It is clear that the same degree of redressive value (V($\mathbf{A}$ B_O)) may be too much, too little or just right depending on the value of the face threat present. Of course, this leaves open the question of how face threat is determined. This aspect of our implementation of the Brown and Levinson theory will be discussed next, followed by a discussion of how we assess redressive actions and their values within this framework.

Algorithm Implementation and Test Cases

In work funded by a DARPA Small Business Innovation Research grant, we have recently completed the initial development of an "Etiquette EngineTM" (EE)—an algorithm based on Brown and Levinson's work as described above and have demonstrated its capability to provide expected politeness assessments both in controlled tests involving project team members and now also in open surveys involving university students unaware of our model. Our approach and results will be described in this section.

The EE Algorithm

We have implemented a version of the Brown and Levinson equation to use as a predictive model of the believability of the redressive actions of an actor (whether human or non-player character (NPC) in a game or simulation environment) as it appears to a human observer, with perceived aspects of context (D, P, R and C). Actors which do not exhibit the expected degree of polite redress (either by being over- or under-polite) are expected to be seen as either unbelievable or to invite rethinking of what was previously understood about the D,P,R and C of the context. For example, if a private bursts in on his captain and issues a bald directive ("Get your coat on") without any redress, an observer might well assume that the degree of imposition (R) is less than might otherwise be the case because the private was charged with giving such instructions, or that the familiarity between them warranted it. Otherwise (and especially in a simulated environment), the observer might simply believe that the private was behaving "unbelievably."

The algorithm we have created operationalizes the equations described above. Expanding on the Brown and Levinson equation, our implementation uses weights on each component to allow the valuing D, P, and R differently (a factor we suspect may underlie some cultural differences), resulting in the equation below:

$$W_x = [w_1 \cdot D(S,H)] + [w_2 \cdot P(H,S)] + [w_3 \cdot R_x] + C(S)$$

Each Observer adds his/her own interpretations of the context. For example, D(S,H) could be expanded to [B_H:w1· B_H:D(S,H)], representing Hearer's belief about the degree of social distance and the Hearer's belief about the appropriate weight for the social distance term. We use Speaker and Observer (who could also be a Speaker or Hearar) belief similarly (B_S: and B_O:). This results in the following for an Observer O:

$$B_O{:}W_x = \{[B_O{:}w_1 \cdot B_O{:}D(S,H)] + [B_O{:}w_2 \cdot B_O{:}P(H,S)] + [B_O{:}w_3 \cdot B_O{:}R_x]\} + B_O{:}C(S)$$

As noted above, the assumption of this model is that the perception of an interaction in context will be determined by the degree to which the weightiness of the face threat is compensated for, or "redressed". Therefore, the value of W_x should be balanced by the "value" (V) of a set of redressive actions used in the interaction x (A_x) if the interaction is to appear "normal" or believable or without ulterior motive. In other words, we expect the value of the redressive strategies the speaker uses to equal or balance the value of the face threat s/he produces. We express this as a difference to give us an "incredibility" or "imbalance" metric which also serves as a perceived politeness metric:

$$B_O{:}I_x = \quad B_O{:}V(A_x) - B_O{:}W_x$$

In order to use this metric to evaluate the imbalance between expected and observed levels of politeness, we must operationalize the various parameters. Space does not permit a detailed presentation of our method for accomplishing this, but we will summarize the basic approach below.

Operationalizing EE Terms

To operationalize and quantify the Brown and Levinson model described above, we first developed scalar values for the politeness parameters P,D and R. These scales were initially for basic American culture, but we have since experimented with representing Pashtu culture in a similar approach with reasonable success. We represented the variables, as well as various parties' perceptions of them, as continuous scalar values ranging from negative to positive infinity. The value of 0 is the "balance point"--a nominal or equal value for each scale, while positive values indicate that the parameter is increased (and contributes to an increasingly "weighty" or potent FTA); negative values indicate that it is decreased (and is building up the Hearer rather than threatening him or her). For Power Difference of the Hearer over the Speaker (P(H,S)), for example, a value of 0 means that the power of the Hearer and the Speaker are equal, that they are (exact) peers. Values greater than 0 indicate that the Hearer (H) has increasingly greater power relative to S and, therefore, that face threat increases whenever S addresses H. Similar scales were developed for D(S,H) and Rx. The character term (C) was represented as a simple value added or subtracted from the W_x sum.

Next, we developed numerical valuations for various redressive behaviors based on the guidelines provided by Brown and Levinson as depicted in Figure 1 above. Ranges of values for the broad classes of strategies were defined as follows, with individual strategies within each class being assigned a value within the designated range:

Table 1. Two sample vignettes.

Vignette 1 —*High Face Threat, High Redress, High Believability*
A low ranking soldier (i.e., a corporal, as indicated by uniform insignia) walks into the Mayor's office and the Mayor motions him quickly to a seat. The soldier takes off his hat and sits down, waiting while the Mayor continues to write something. The Mayor finishes up writing, puts down his pen and looks up at the soldier expectantly. The soldier then says, "I'm sorry to interrupt you work, Mayor Fredrickson, but my name is Corporal Jones and I've been put in charge of your escort to the event tonight. I was wondering if it would be possible for you to let me know where I can meet your wife so that I can get her there on time?"

Vignette 2—*High Face Threat, Low Redress, Low Believability*: As for vignette 1 above except that the soldier acts and speaks differently. Here, he interrupts the mayor while he is speaking, perhaps by putting a hand on his shoulder, and says loudly, "Tell me where I can meet your wife?"

- The value of the use of an individual positive redressive strategy (see Brown and Levinson, p. 102, Figure 3, for a list) will provide from 1 to 40 "units" of redress.
- The value of the use of an individual negative redressive strategy (see Brown and Levinson, p. 131, Figure 4, for a list) will provide from 20 to 60 units of redress.

Within the range defined above, a specific score was assigned to individual instances of redress which fell into the category, as will be illustrated below.

The effects of multiple redressive strategies were scored as simply additive. We understand that this is a simplification, and that the efficacy of added redressive behaviors probably falls off, eventually becoming simply irritating. This means that the value V of a set of N redressive actions A contained in interaction x is:

$$V(A_x) = V(A_1) + V(A_2) + \dots V(A_N)$$

Evaluation Test Cases

This approach was tested in a series of sample social interaction vignettes crafted to represent (according to our American cultural intuitions) either normal/balanced politeness, unbelievable over-politeness or unbelievable rudeness. Our goal was to determine if the equation and scoring techniques we had developed would track our intuitions. The level of face threat and redress were varied over this set of vignettes so that high face threat situations were paired with high levels of redress (roughly balanced) as well as low levels of redress (highly imbalanced and rude). Similarly, very low levels of face threat were paired with very high levels of redress (over-polite) and with low levels of redress (balanced). Examples of two such vignettes are illustrated in Table 1.

Table 2. Scoring of Redressive Behaviors used in Vignette 1.

Action and Interpretation	Score
1. The soldier waits until the mayor is finished and invites him to speak. This seems to be a very explicit form of negative politeness (putting the other's interests first) and, especially in this instance where the H was not actively engaged in another conversation, seems very potent.	60
2. The soldier takes off his hat. This is a sign of deference, which is in turn a fairly potent negative politeness strategy.	50
3. The soldier apologizes for interrupting. This is also a negative politeness strategy, though arguably a less potent one (though that may be highly mitigated by facial expressions and body language).	30
4. The soldier uses an honorific. Moderately potent negative politeness strategy.	40
5. The soldier poses the FTA as a question. Common negative politeness strategy.	20
6. The soldier offers an explanation/reason for needing the information. Positive politeness strategy.	35
7. The soldier appeals to the Mayor's (H's) interests. Positive politeness strategy Powerful in this context.	30
8. The soldier is hesitant and skeptical about compliance. A common but reasonably potent negative politeness strategy.	30
TOTAL	295

Evaluation 1—Trained Rater Correlations

Each of the eight vignettes was then assessed using the operational scoring tables we had created for both the situational context parameters (P, D, R, and C) and for the values of the individual redressive actions used. For example, for the first vignette the imbalance evaluation proceeded as follows:

- The corporal (as S) has lower power than the mayor by a fairly large degree. Their "power difference" is fairly large—perhaps slightly larger than our anchor point of 100, yet less than the anchor point of 1000. We scored that as $P(H,S) = 300$ (and, since there were no cultural differences or speaker or observer misperceptions $B_O:P(H,S) = 300$).

- There is no particular familiarity between the two individuals in this vignette, but social distance is not extreme either. They are from slightly different "cultures" (military vs. civilian infrastructures) and show no evidence of prior relationship, but they are engaged in a common endeavor. The social distance between them is probably only slightly higher than 0. Thus, $D(S,H) = 3$.

- The imposition of this request could be somewhat large. To ask after the location of one's wife so as to pick her up is comparatively threatening, though the fact that this is in the mayor's service should mitigate this imposition (as the corporal reminds him). The raw imposition is a short answer required from the mayor, characteristic of our level 10, so we assigned it: $R_x = 10$.

- Since we have provided no reason to believe that the character of the corporal is anything other than nominal, we will assume that $C=0$.

This gives us a value of $B_O:W_x$ as (using the left hand portion of the equation) as: $3 + 300 + 10 = 313$.

For the value of the redress applied $V(A_x)$ we scored the set of redressive actions in Table 2.

Thus, the imbalance score for this vignette, as calculated by our equation, would be: $295 - 313 = -18$. Since this vignette was intended to convey both high face threat and high redress and, thus, to be balanced and believable, this score seems to be about right, falling very near zero. For the second vignette, by contrast, we have a high degree of face threat with virtually no redressive actions. This is unexpected and should be perceived as very rude. This scenario should have a score much less than 0 on our imbalance metric—indicating that there is substantial unredressed threat. This vignette had the same W_x attributes as Vignette 1 and was was scored as having only 40 points of redress, thus giving an I_x score of $40-313 = -273$. This is a strongly negative score—as we expected for an interaction intended to be perceived as rude.
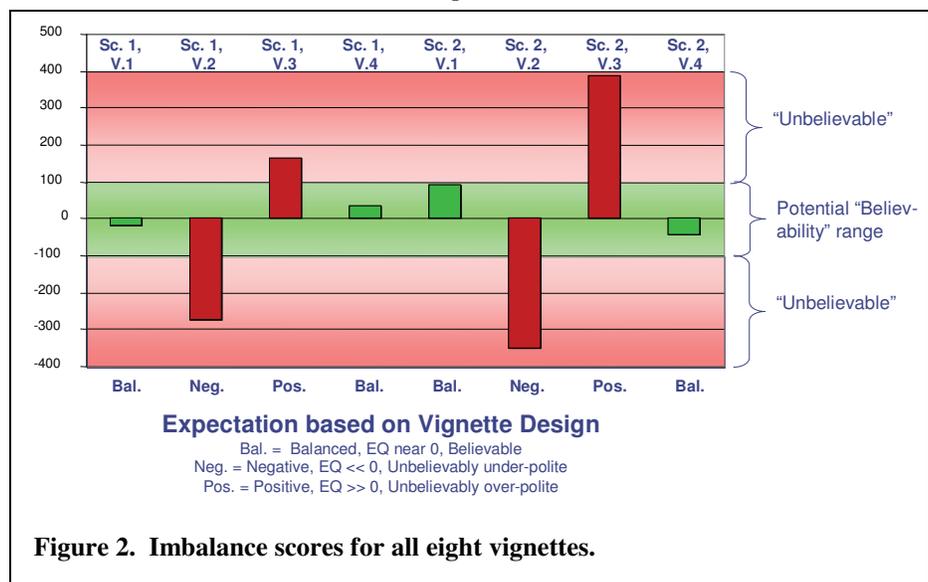

Figure 2. Imbalance scores for all eight vignettes.

Table 3. Correlations between mean scores provided by trained team member using the EE algorithm and untrained college students.

Variable Asked About	Correlation	Significant?
Power Relationship (= P)	.867	--
Familiarity (= D)	.881	--
Imposition (= R_x)	.766	--
Overall Politeness (= I_x)	.892	yes

An evaluation similar to that described above was carried out for a total of eight vignettes and the quantitative algorithm tracked predictions for rude, polite or nominal perceived etiquette levels very closely. As shown in Figure 2, all vignettes that were intended as "nominal" interactions (that is, using about the amount of redress as would be expected in American culture for the amount of redress offered) scored within +/- 100 points of zero. All vignettes that were expected to be seen as over-polite scored well higher than 100 points; while all that were expected to be seen as overly rude scored substantially less than -100 points.

While the above example was based on one individual's scoring assessments (Dr. Miller's), we have since replicated this work with two other "raters" following a brief training session. Each rater was a member of the project team and was generally familiar with the Brown and Levinson model and our use of it, but not of the specific scores that Dr. Miller had produced. Each rater scored the vignettes and the results of the three raters' scores were then statistically compared. The top-level imbalance metric (I_x) showed a Robinson's A correlation of .931 among the three raters across the 8 vignettes, and the two major sub-factors (Face Threat Weight—W_x and Redress Value—$V(A_x)$) showed Robinson's A correlations of .950 and .863 respectively. These values are all well above traditional correlation thresholds of .7 or .8 for multiple judge rating correlations. Thus, this study lends weight to the claim that we have identified a reliable method of scoring the degree of politeness in social discourse—at least in American cultural settings.

Evaluation 2--Untrained Rater Correlations

Subsequent to the evaluation involving trained team members, we conducted an experiment wherein 22 American college students, unaware of our theory or model, also rated various aspects of the vignettes. The same eight vignettes were used. Participants reviewed a "backstory" describing the relationships between participants in the vignette and then answered a series of questions about the relationships between actors in the vignette. They then read the specific, verbal interaction between the actors, and then answered a subsequent set of questions about their perceptions of the actors, their relationships, the degree of politeness used and whether or not they regarded the interaction as normal, rude or overly polite.

Correlations between participants' ratings of the relevant model parameters and our own ratings remained very high, as shown in Table 3. Pearson correlation coefficient scores are reported for a comparison of the project team's rating with the untrained participants' mean rating. Significance was assessed at the $p<.01$ level (two-tailed). Although these correlations are very high in general, the small number of cases evaluated (only four means were assessed for the P, D, and R ratings, since the relationships were identical in pairs of vignettes) kept many of the relationships statistical significance.

Participants were also asked whether they changed beliefs about the P and D values after they had seen the utterance used in the vignette. Since some vignettes used levels of politeness rated nominal by our scoring algorithm while others used either unexpectedly high or low levels of politeness, we hypothesized that if our model were correct, then more participants would be willing to change their ratings after seeing the vignettes with "off-nominal" politeness rather than those with nominal politeness.
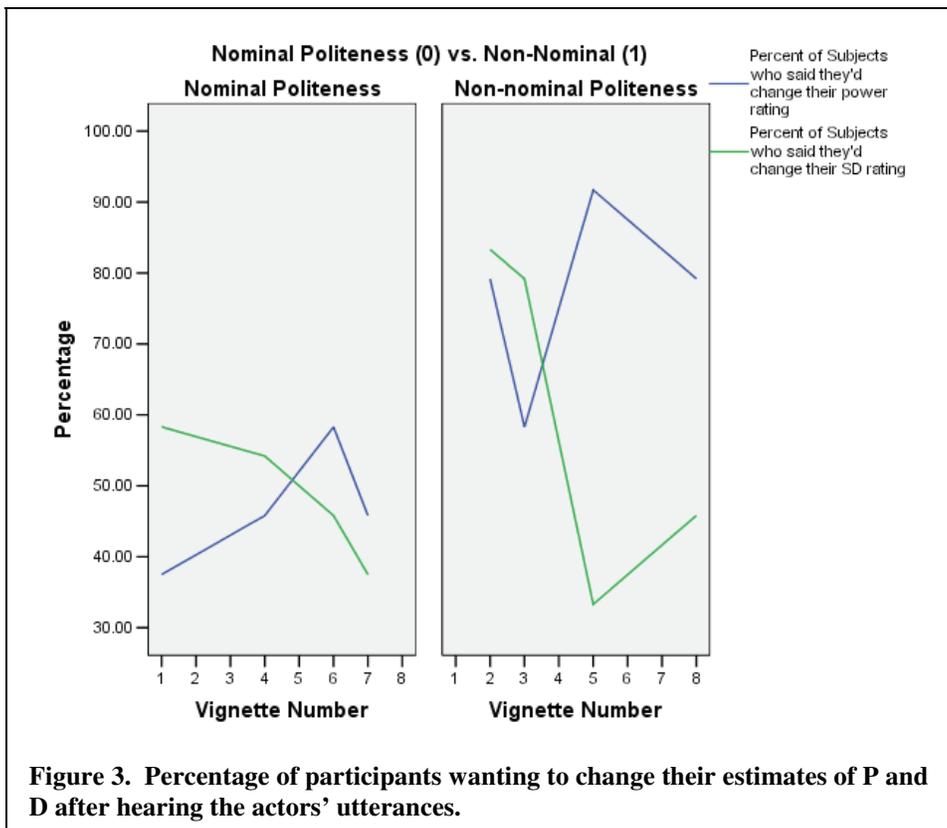
As can be seen in Figure 3, a higher percentage of participants generally reported willingness to change their ratings in response to the off-nominal vignettes rather than the nominal ones, though this effect was more pronounced for Power Distance than for Familiarity (Social Distance) ratings. Table 4 reports the overall percentage of participants, taken over the four vignettes in each condition (nominal vs. off-nominal politeness) who reported wanting to change their ratings of either Power Difference or Social Distance after reading or viewing the actors' behaviors. A paired-samples t-test on the mean values for the four nominal vs.

	Nominal Vignettes	Off-Nominal Vignettes
Power Difference	46.9%	77.1%
Social Distance	48.9%	60.4%

Table 4. Proportion of participants wanting to change their P and D ratings averaged over vignette types.

four off-nominal vignettes showed that significantly more participants wanted to change their estimate of Power Difference after reading/viewing the Off-nominal behaviors than the Nominal ones ($t=-4.85$, $df=3$, $p<.05$). A similar test for the Social Distance parameter was not significant ($t=-1.186$, $df=3$, $p>.2$).

In general, these data support our model and its claims that unexpected (i.e., off-nominal) amounts of redress prompt people to reinterpret their beliefs about context—specifically, their beliefs about the P, D (and potentially, R and C) parameters. In this study, at least, participants proved more willing to review their perceptions of Power

Figure 3. Percentage of participants wanting to change their estimates of P and D after hearing the actors' utterances.

implications) over which the model's parameters are scored. This has the practical implication that the general social interaction reasoning of an automated system can be effectively modularized, and, thus, large savings in simulation code development can be realized. Furthermore, basic game storylines or training modules and even specific characters can be easily transposed from one cultural milieu to another—enabling the village priest who the player had to interact with to get intelligence information in a Kosovo training game to take on the culture-specific behaviors and reactions of an imam in an Iraq training game. In each case, new knowledge bases of culture-specific politeness behaviors would need to be developed (and, of course, checked for accuracy) for each new game, but the core game storyline(s) and character roles, general actions, motivations, capabilities, etc., could remain unchanged.

relationships than Familiarity. This may be a function of the marked power relationships in the vignettes (involving, as they did, soldiers and civilians) or it may reflect a more general tendency among Americans to seek explanations for politeness variations in power dimensions first.

Conclusions and Future Work

An ability to score the believability of the social interaction behaviors of actors in a cultural and social context is important because it allows for quantitative reasoning (and, ultimately, for machine aids and predictions) about how social interaction "moves" will be perceived. Thus, this model holds the potential to equip a wide range of simulated characters from different cultures with the ability to evaluate human behaviors given what it knows about P,D,R and C and then to reactively take offense or take advantage. Similarly, it can equip machine systems with the ability to determine appropriate behaviors in order to further ends such as trust, usability and user acceptance.

The use of Brown and Levinson's model and theory in a module for reasoning about social interaction behaviors ensures universal reasoning about and scoring of abstract politeness "moves". Any such module will need to be equipped with culture-specific knowledge bases, however, to enable reasoning from the observable behaviors in a culture (e.g., pursed lips or a rigid hand-to-eyebrow salute) to the abstract etiquette "moves" (and therefore, politeness

In work reported elsewhere (Miller, et al., 2007), we have reported on our work integrating this implemented algorithm into a language training game (the Tactical Language Training System developed by USC's CARTE Labs—cf. Johnson, Vilhjalmsson, and Marsella, 2005). This work has demonstrated the capability of our approach to inform both the perceptions and the reactions of simulated characters—and to do so with less software development time than traditional scripting approaches. Moreover, it has also demonstrated the ability of our approach to provide at least reasonable knowledge and use of politeness levels in a culture different from American English, (specifically, the Pashtu language spoken along the Afghanistan/Pakistan border), though we haven't performed as thorough an evaluation of those results as of the Amercian cultural results presented above.

While etiquette and politeness are far from the only aspects of culture that should be modeled in computational tools and approaches, they are a pervasive aspect of virtually all interactions that matter in cross-cultural collaboration and interaction. In this work, thanks to the basic model developed by Brown and Levinson, it is proving ammenable to quantitative, computational modeling. Furthermore, the resulting models are providing predictions that correlate well with both trained and naïve users of our modeling framework, at least for American cultural sensibilities.

Acknowledgements

References

Brown, P. & Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge,UK.; Cambridge Univ. Press.

Cassell, J. and T. Bickmore. (2003). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction. 13*(1): 89-132.

Dennet, D., (1989). *The Intentional Stance*, Cambridge, MA; MIT Press.

Goffman, E. (1967). I*nteraction Ritual: Essays on Face to Face Behavior.* Garden City; New York.

Hofstede, G. (2001). *Cultures and Consequences*, 2nd Ed. Thousand Oaks, CA; Sage Publications.

Johnson, L. and Rizzo, P. (2004). Politeness in Tutoring Dialogs: "Run the Factory, That's What I'd Do". *Intelligent Tutoring Systems*, 67-76.

Johnson, L., Vilhjalmsson, H. and Marsella, M. (2005). Serious Games for Language Learning: How Much Game, How Much AI?. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education* . July 18-22, Amsterdam, The Netherlands.

Miller, C., Wu, P., Funk, H., Wilson, P. and Johnson, L. (2007). A Computational Approach to Etiquette and Politeness: An "Etiquette Engine™" for Cultural Interaction Training. In *Proceedings of BRIMS 2007*. Norfolk, VA; March 26-29.

Nisbett, R.E. (2003). *The geography of thought: How Asians and Westerners think differently…and why*. NY: Free Press.
.