Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness

Eleonore Lumer (eleonore.lumer@uni-bielefeld.de)
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies
Bielefeld University, Bielefeld, Germany

Hendrik Buschmeier (hbuschme@uni-bielefeld.de)
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies
Bielefeld University, Bielefeld, Germany

Abstract

Politeness is a social linguistic phenomenon. Modeling polite language production and understanding is difficult, as it may contradict conversational maxims and is shaped by extralinguistic social influences, such as the speaker-hearer relationship. This paper extends Yoon et al.'s (2016) Rational Speech Act-based model of politeness by mapping speaker-hearer relationship influences to the utility weights of the model and instantiates it in German. Three online experiments, for empirical analysis and collection of behavioural data for model training and evaluation, are presented. These confirm the influence of the speaker-hearer relations on indirect politeness. Furthermore, two versions of the model are trained and evaluated to find out which part of the model is better suited for the integration of social influences. Overall, both model versions yielded similar results and were able to predict the meaning of polite speech acts.

Keywords: computational modeling; politeness; rational speech act; speaker-hearer relations

Introduction

Politeness is a widely researched phenomenon, in which not only different linguistic research fields such as pragmatics, syntax, sociolinguistics and psycholinguistics are interested, but also other cognitive and social sciences (Brown, 2015; Watts, 2003; Kasper, 1990; Fraser, 1990). In this paper, politeness is considered as a set of face saving strategies, meaning that it is used to enhance or preserve the public self-image of a listener or speaker (Brown and Levinson, 1987).

One aspect that makes politeness an interesting phenomenon for formal modeling is that its use often stands in opposition to conversational principles, such as the Gricean Maximes (Grice, 1975). This can for example occur for indirect ("off-record") politeness (Brown and Levinson, 1987), where speakers choose to say something indirectly in order to save the face of the listener – even though the informativeness and truthfulness of the utterance might suffer. In a recent cognitive modeling approach based on the *Rational Speech Act* (RSA) model (Yoon et al., 2016, 2017, 2020) this balancing of competing goals is used to model the understanding and production of politeness. Previous research on politeness has identified social influences on the choice of politeness strategy (Brown and Levinson, 1987; Holtgraves and Bonnefon, 2017). The RSA-based politeness model, however, did not consider "mapping from social situations into utility weights and communicated social weight" (Yoon et al., 2020, p. 80).

This paper focuses on a specific social influence on politeness, the speaker-hearer relation, and extends (a German instantiation of) the RSA-based model of Yoon et al. (2016) with it. As an example, an influence of the speaker-hearer relation on politeness could be that one feels more free to tell an unpleasant truth to a friend than to one's boss. Two versions of the model were implemented, one ($pRRSA_c$) maps the speaker-hearer relations to the speaker's goal, the other ($pRRSA_f$) to the scaling parameter of politeness.

This paper has two goals: to analyze which parameter is more adequate for mapping the social influence (comparing the two model implementations), and to find further empirical evidence for the influence of speaker-hearer relations on indirect politeness. The data collected across three experiments is used for training and testing the models.

Social influences on politeness

The theory most often used as the basis for current politeness research and modeling is Brown and Levinson (1987)'s politeness theory (Leech, 2014; Watts, 2003; Brown, 2015). This theory is based on Goffman's (1955) concept of "face" and proposes the use of different politeness strategies based on the degree of face threat of a conversational act. According to Brown and Levinson (1987), the degree of face threat is influenced by three factors: the *power* of the listener over the speaker, the *social distance* between speaker and listener, and the (culturally dependent) *rank of imposition*. For example, an act that has a low degree of face threat will likely result in a more direct strategy ("on-record"), whereas an act with a high degree of face threatening potential will likely result in an indirect politeness strategy ("off-record").

Several studies found influences of power on politeness: Leichty and Applegate (1991) studying the situational influences, Ambady et al. (1996) with a cross-cultural study, or Danescu-Niculescu-Mizil et al. (2013) with a corpus analysis; see Leech (2014) and Holtgraves and Bonnefon (2017) for further pointers into the literature. The influence of social distance on politeness, however, appears to be less clear (Holtgraves and Bonnefon, 2017). One suggested explanation was Brown and Levinson (1987)'s definition of social distance – based on the frequency of interaction – and it was suggested to further include the likeability of a person in the distance evaluation (Vergis and Terkourafi, 2015). Other research on the social influences on politeness found the presence of third parties (Leech, 2014; Watts, 2003) and gender (Holtgraves and Bonnefon, 2017; Kasper, 1990; Carli, 1999) to play a role.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

For the latter, it was, however, argued that the influence could also be due to an underlying perceived power and distance difference (Holtgraves and Bonnefon, 2017). As power and distance are fundamental for politeness understanding and production, we consider them for our implementation of social influence in a polite *Rational Speech Act* model.

RSA-based modeling of politeness

The *Rational Speech Act* model (RSA) is a probabilistic approach to modeling pragmatics (Franke and Jäger, 2016), based on game theory (Frank and Goodman, 2014) and Bayesian statistics. RSA has been used to model various linguistic phenomena (Goodman and Frank, 2016), one of them being politeness (Yoon et al., 2016, 2017, 2020). This paper is based on the RSA model of politeness as described in Yoon et al. (2016), which proposed the consideration of a *social utility* (U_{social}) in addition to an *epistemic utility* ($U_{\text{epistemic}}$) of a communicative act, thereby modeling the conflict between 'being nice' and 'being informative' when producing an utterance. In the explanation that follows, we will only focus on these two utilities and their interplay (see Yoon et al. (2016) and Scontras et al. (2021) for model details).

Conceptually, RSA is a recursive model of a pragmatic listener (P_{L_1}) doing Bayesian reasoning about a speaker (S_1) in order to infer the state of the world (s), which the speech act (u) refers to. RSA models the epistemic utility

$$U_{\text{epistemic}}(u; s) = \ln(P_{L_0}(s \mid u))$$

which represents the informativeness of the model of the literal listener (L_0), by mapping u to a state s by considering the literal meaning and the prior distribution of the states: $P_{L_0}(s \mid u) \propto u(s) \cdot P(s)$. The data for the states' prior distributions in L_0 is collected in experiment 1 of this paper. When modeling politeness in RSA, Yoon et al. (2016) introduced a social utility

$$U_{\text{social}}(u; s) = \mathbb{E}_{P_{L_0}(s|u)}[V(s)]$$

to model the competing goal of being nice. This utility uses a value function $V(s) = \alpha \cdot s$ that manipulates the degree of politeness. Simply put, the social utility is responsible for making the speaker choose a more positive speech act than the literal meaning when $\alpha > 1$. Both utilities (and hence the competing goals of a speaker) are represented in the overall utility function

$$U(u; s; \phi) = \phi \cdot U_{\text{epistemic}}(u; s) + (1 - \phi) \cdot U_{\text{social}}(u; s)$$

where $\phi \in [0, 1]$ is the goal weight defining how nice (indirect) or informative (direct) the speaker intends to be.

An updated version of this model (Yoon et al., 2020) adds a third goal (and utility) that considers the *self-presentational* goal of a speaker. This allows the speaker to not only consider the degree of informativeness and niceness, but also that she wants to be seen as being considerate. In this paper we concentrate on the original model by Yoon et al. (2016), as it introduced the concepts of social utility weighing that is the relevant factor for our approach.

In the "polite RSA" model by Yoon et al. (2016, 2017), the speaker strategy does not consider social influences. We hence introduce two altered versions of it in the form of "polite Relational RSA" (pRRSA) models in order to implement speaker-hearer relation influences. In both of these versions the model stays the same and only one parameter in each version is adapted by being fitted to the different relationships. In the first model, pRRSA_c, the speaker-hearer relation influences *the choice of a speaker's strategy* (informative vs. nice). This is implemented by setting the value of ϕ depending on the relationship. The approach suggests, that listeners are aware of their relationship to a speaker and evaluate how polite or direct the speaker might choose to be. From the perspective of speech production, pRRSA_c could be interpreted to consider the relationship influence at the "conceptualisation" stage (Levelt, 1989). In contrast to this, in the second model, pRRSA_f, the speaker-hearer relation influences the *choice of the degree of politeness*. This is implemented by setting a different pretrained value α based on the relationship, which then modifies the degree of divergence to the literal meaning in the value function $V(s)$. From the perspective of speech production, pRRSA_f could be interpreted to consider the relationship influence at the "formulation" stage (Levelt, 1989). To the best of our knowledge this is the first work implementing relationship influences in RSA and also the first test of polite RSA in another language.

Expectations and hypotheses

For the three experiments and for the models we had the following expectations. For experiment 1 we expected similar literal meaning evaluations for the translated German target words as those in Yoon et al. (2016). Experiment 2 collects evaluations of indirectness for speaker strategies based on the speaker-hearer relation, which will be used as an empirical comparison to the optimized ϕ values in the pRRSA_c model. Experiment 3 will collect data for model training and will additionally be used to analyze the influence of speaker-hearer relations on the evaluations of the meaning of target words.

The speaker-hearer relations used in experiments 2 and 3 are based on previous research on power and distance that also uses relation designations such as *close friend* to analyze the influence of assumed speaker-hearer relations on politeness (Holtgraves and Yang, 1990; Kasper, 1990; Vergis and Terkourafi, 2015; Watts, 2003). Relationships that do not differ in authority or status, e.g., *friends* or *colleagues*, are regarded as having similar power (Brown and Levinson, 1987; Holtgraves and Yang, 1990). While relations including roles with authority and higher status, such as a *boss*, are seen as having higher power. The distance is manipulated with adjectives, suggesting more distance (*dreaded*, *distant*) or less distance (*close*, *easy-going*) between interlocutors. Due to the speaker-hearer relationship, the degree of face threat for the same utterances differ (Brown and Levinson, 1987). Relationships considered to have equal power and distance, such as *close friends*, are expected to cause speakers to use

more direct politeness strategies. Previous research found that when coming from a person with lower distance and power, direct and potentially face threatening utterances are accepted (Brown and Levinson, 1987; Holtgraves and Bonnefon, 2017; Gupta et al., 2007). In contrast to this, when a listener has high power over a speaker, e.g., a *dreaded boss*, the choice of a speaker's strategy is expected to be indirect as the face threat is high (Holtgraves and Yang, 1990). In experiment 3 we therefore expect that indirectness results in more negative state evaluations for positive target words. Furthermore, we expect negative target words to differ less in their evaluations between relationships due to a floor effect (more negative evaluation is not possible). Finally, we expect our modeling approach to achieve a similar performance as the model presented in Yoon et al. (2016), as the main difference lies in the data and not the model structure. Specificly, our hypotheses are:

H-1: Different speaker-hearer relations (with different power and distance) account for higher or lower face threats.

H-2: Higher face threats trigger indirect politeness.

H-3: Degree of positivity of the target word influences the evaluated degree of face threat.

H-4: Polite RSA (Yoon et al., 2016) can be instantiated with German language data.

Experiment 1: Literal semantics

Experiment 1 collects data for the model's literal listener (L_0). It is based on the literal semantics experiment by Yoon et al. (2016) – but with target words in German.

Method

Participants Participants were mostly recruited via text message and did not receive compensation. As in Yoon et al. (2016), a relatively small number of 32 participants, all German native speakers, took part. Of these, most were university students (75%) and female (68.75%). 87.5% of participants were aged 22–30.

Material and design We translated Yoon et al. (2016)'s target words ("amazing", "good", "okay", "bad", "terrible"), representing different judgements of a situation, to German: "großartig", "gut", "okay", "schlecht", "schrecklich".

Procedure The task for participants in this experiment was to judge whether they think a target word matches a certain state ('meaning') represented on a five-point heart-shaped scale (♡–♡♡♡♡♡). Each participant had to rate all target words in combination with all possible states. Each item was randomly presented in one out of five short context scenarios concerning the evaluation of a hobby, e.g.: *Lisa baked a cake for Sue. Sue tastes it and rates it with three hearts* (♡♡♡). *Does Sue think that the cake is* "schrecklich"? Participants had to answer either "yes" or "no". To avoid gender effects on politeness, all characters used in the scenarios were female (Kasper, 1990).

Results

Table 1 shows the proportion of participants' acceptance of target words given a state (from one to five hearts). Results were collapsed across contexts for the evaluation. For each target word a different distribution over the states can be observed.

Table 1: Acceptance ("yes") proportions of target word–state matching judgements by participants in experiment 1.

Target word		State				
DE (ours)	EN (Yoon)	1 ♡	2 ♡	3 ♡	4 ♡	5 ♡
großartig	amazing	0	0.03	0.06	0.16	1
gut	good	0.03	0.03	0.66	1	0.90
okay	okay	0.06	0.47	0.94	0.66	0.47
schlecht	bad	0.91	0.59	0.06	0	0
schrecklich	terrible	0.84	0.25	0	0	0

We also compared the results to Yoon et al. (2016)'s using English target words. Overall, the state evaluations were similar, the German translations seemed to be perceived to be more negative though (lower state evaluations).

Experiment 2: Evaluation of indirectness

Experiment 2 measures the influence of speaker-hearer relationship and state on the choice of directness of a speaker's strategy. Directness values obtained in this experiment will be compared to the predicted values of ϕ in the pRRSA_c model. The empirical results will also be compared to those from experiment 3 in order to test the relationship influences on the choice of (polite) indirectness.

Method

Participants Compared to the questionnaire of experiment 1, the recruitment for this experiment was extended to social media platforms and the university department's mailing list. Participants could take part in a raffle to win a 20 EUR voucher of their choice. A total of 126 participants were recruited, of which 84 remained after excluding non-native speakers of German as well as participants who failed attention-check questions. Most participants were students (65.5%), and female (61.3%). The average age of all participants was 26.28 (SD = 8.41).

Material and design As in experiment 1, five fictional context scenarios concerning the evaluation of a hobby were used and participants received information on the opinion of the speaker ('state') represented on the five-point heart-shaped scale. Additionally, four speaker-hearer relationships, expected to differ on the dimensions of power and social distance, were used. The role of the listener in a scenario was either a *close friend* ("enge Freundin"; suggesting low power and small distance), a *distant colleague* ("entfernte Kollegin"; low power, large distance), an *easy-going boss* ("lockere Chefin"; high power, small distance), or a *dreaded boss* ("gefürchtete Chefin"; high power, large distance). Again, all fictional characters were female. Each participant received 20 items, combining every possible combination of relationship and state, as well as three attention-check questions.

Procedure Each item first described the speaker-hearer relationship, a random context scenario and the opinion of the

Table 2: Median directness values per relationship and state, from experiment 2 (0/"very direct" – 100/"very indirect").

Relationship	State				
	1 ♡	2 ♡	3 ♡	4 ♡	5 ♡
Close friend	34.5	35.5	23.5	7.5	0
Distant colleague	85.0	73.0	54.0	24.0	8.5
Easy-going boss	69.5	62.0	39.0	16.0	0
Dreaded boss	98.5	87.0	66.5	27.0	13.0

speaker on the heart-shaped scale. The listener then asks for the speaker's opinion on her performance, e.g., *"Lisa asks Sue for her opinion about the cake."* Participants were then asked how the speaker (e.g., Sue) would respond to the question using a slider-based scale from 0 ("very direct"; i.e., on-record) to 100 ("very indirect"; i.e., off-record). These terms were explained to participants based on Brown and Levinson's (1987, pp. 68–69) explanation of on- and off-record strategies.

Results

A one-way analysis of variance (ANOVA) was conducted in order to evaluate speaker-hearer relationship influences. We found statistically significant differences between relations ($F(4,3) = 161.259, p < 0.001$) and states ($F(4,3) = 389.317, p < 0.001$). As can be observed in table 2, for positive states (four or five hearts) participants chose more direct speaker strategies across all relationships. Due to the large range of the response scale (0–100), we consider median values for more direct inspection.

Discussion Generally, the data was rather distributed over the response range, especially for the relationships *close friend* and *easy-going boss*. Tendencies confirming hypotheses H-1 and H-2 can be observed though: Relationships with low power and small distance (*close friend*) result in a more direct strategy than those including a listener with high power and large distance (*dreaded boss*). In line with intuition, participants did not see the need for an indirect speaker-strategy for higher states (i.e., more hearts). Additionally, it can be observed that participants chose a more indirect strategy when the listener had a large distance and low power (*distant colleague*) than when the listener had high power and small distance (*easy-going boss*). This could be interpreted as an indication for a difference in relevance of power and distance for indirectness. This result is surprising given the rather unclear findings regarding the influence of distance on politeness – as opposed to the conclusive evidence for the influence of power – in previous research (Holtgraves and Bonnefon, 2017; Leichty and Applegate, 1991).

Experiment 3: True state inference

Experiment 3 collects data for model training as well as to gain further insights on the influence of relationships and states on indirect politeness interpretation. It is an adaption of Yoon et al. (2016)'s experiment 2.

Table 3: Mean values and standard deviations for each relationship and target word from experiment 3. The two columns on the right show the predicted mean results of the two models.

Relationship	Target word	Exp. 3 Mean (SD)	Model predictions $pRRSA_c$	Model predictions $pRRSA_f$
Close friend	großartig	4.80 (0.46)	4.38	4.26
	gut	3.31 (0.74)	3.54	3.52
	okay	2.38 (0.71)	2.44	2.50
	schlecht	1.12 (0.33)	1.17	1.21
	schrecklich	1.18 (0.73)	1.04	1.07
Distant colleague	großartig	4.31 (0.78)	3.93	3.87
	gut	2.91 (0.71)	3.09	2.25
	okay	2.04 (0.65)	2.21	2.15
	schlecht	1.26 (0.61)	1.06	1.05
	schrecklich	1.04 (0.24)	1.02	1.01
Easy-going boss	großartig	4.46 (0.75)	4.3	4.1
	gut	3.27 (0.72)	3.50	3.36
	okay	2.30 (0.67)	2.41	2.37
	schlecht	1.25 (0.49)	1.14	1.12
	schrecklich	1.07 (0.37)	1.03	1.04
Dreaded boss	großartig	3.69 (0.98)	3.62	3.60
	gut	2.59 (0.87)	2.37	2.25
	okay	1.67 (0.73)	1.89	1.77
	schlecht	1.19 (0.69)	1.04	1.01
	schrecklich	1.11 (0.35)	1.02	1.00

Method

Participants This was part two of an online study that combined experiments 2 and 3, participants are thus the same.

Material and Design Instead of using speakers goals as the independent variable (Yoon et al., 2016), we used the speaker-hearer relationship. This substitution was done, as it was considered difficult for a listener to know or guess the goal of the speaker, without considering other influences. Hence, we implemented speaker-hearer relations as a social influence on the degree of utility weighing. Each participant received one out of five possible context scenarios for every combination of relationship and target word.

Procedure Each item first described the relationship (from exp. 2) between the characters (e.g., *"Lisa is Sue's easy-going boss"*), followed by a short context scenario that included the listener's demand for feedback (e.g., *"Lisa baked a cake and wanted to know how Sue liked it"*). The speaker's utterance was given containing a target word (from exp. 1, e.g., *"Sue says: 'It was* gut*'"*. Afterwards, participants were asked to answer on a five-point heart-shaped scale what they believed the speaker (e.g., Sue) actually thought (state).

Results

For the evaluation, the results were again collapsed across contexts. This experiment collected the influences of speaker-hearer relations and target word on the state evaluation. Means were compared and analyzed with a one-way ANOVA and Tukey post-hoc tests (see supplementary material for details).

Table 3 shows the mean state evaluation and standard deviation results for each target word and relationship. Overall,

the target words "schlecht" and "schrecklich" were perceived similarly. It can further be observed that the standard deviations for the *dreaded boss* were higher than those for the other relationship conditions. The results of the one-way ANOVA showed a statistically significant difference between relationships ($F(3,4) = 48.58, p < 0.001$) and the different target words ($F(3,4) = 1416.91, p < 0.001$). Further, a statistically significant interaction of the two variables with a small effect size ($n^2 = 0.015$) was found. Tukey post-hoc tests were conducted in order to analyze which relationships and target words differed. For almost all relationship pairs the results differed statistically significantly, except for the comparison of the results for the *close friend* and *easy-going boss*.The relationships that triggered the most different results were the *close friend* and *dreaded boss*. Apart from the target words "schlecht" and "schrecklich", all target words differed statistically significantly from each other. Overall, the evaluations for the more negative target words did not differ greatly between relationships, while the target word "okay" differed the most.

Discussion The finding, that relationships with higher power and larger distance triggered lower state evaluations, support the results from experiment 2. This suggests that participants assumed the speaker to be polite (indirect) by choosing a more positive target word than what was actually meant. This is in line with the results from experiment 2 and hypotheses H-1 and H-2 can therefore be accepted.

Further, as expected, more negative target words did not differ greatly between relationships. This can be explained with the interpretation that already negative target words might either be assumed to be honest, or participants were not able to choose a worse interpretation or say that it was overall inappropriate (e.g., saying to one's boss that something she did was terrible). Hypothesis H-3 can however be accepted.

Modeling

The model implementation with the training of the parameters and the prediction is based on scripts by Yoon et al. (2017, as materials of Yoon et al. (2016) were not available). The formal description of the models is equal to that of Yoon et al. (2016), existing parameters remain the same but with individual weighing parameters ϕ and α for each relationship.

Model training

The same steps were conducted for both model versions. Parameters were calculated using Markov chain Monte Carlo (MCMC) methods (Kruschke, 2014; Yoon et al., 2016, 2017) and then used to predict the states for each target word and speaker-hearer relationship. The data collected in experiment 3 was split 80/20 into a training set for MCMC and a test set for the evaluation. The MAP values from the results of experiment 1 were used for the literal listener model. For the parameter optimization of ϕ (one for each relationship in pRRSA_c), α (one for each relationship in pRRSA_f) and λ, two MCMC chains with 80.000 iterations were calculated for each model. Uninformative priors were used for all variables as prior probabilities: $\phi \sim U(0,1)$ (four ϕ variables in pRRSA_c), $\alpha \sim U(0,5)$ (four α priors in pRRSA_f), and $\lambda \sim U(0,20)$. Table 4 shows the MAP results and highest probability density intervals of the parameter optimizations.

Table 4: Maximum a posteriori (MAP) estimates and 95% highest probability density intervals (HDI) for the parameters resulting from the MCMC optimization. The rightmost column shows the values chosen for the model predictions.

Model	Variable	MAP [95% HDI]	Values
pRRSA_c	α	0.61 [0.45, 1.14]	1.28
	λ	4.59 [3.11, 5.71]	3.4
	Close friend | ϕ	0.45 [0.34, 0.59]	0.58
	Distant colleague | ϕ	0.31 [0.25, 0.47]	0.47
	Easy-going boss | ϕ	0.37 [0.32, 0.56]	0.56
	Dreaded boss | ϕ	0.22 [0.15, 0.33]	0.34
pRRSA_f	ϕ	0.69 [0.6, 0.75]	0.66
	λ	2.25 [1.94, 2.63]	2.37
	Close friend | α	1.88 [1.24,2.54]	1.86
	Distant colleagues | α	2.9 [1.94, 3.69]	3.04
	Easy-going boss | α	1.97 [1.36, 2.71]	2.33
	Dreaded boss | α	4.63 [3.01, 5]	4.74

Table 5: R^2, top1- and top2-accuracy values for each relationship for the pRRSA_c model predictions.

Relationship	R^2	Top1-Acc.	Top2-Acc.
Close friends	0.79	0.72	0.95
Easy-going boss	0.74	0.64	0.90
Distant colleagues	0.78	0.75	0.87
Dreaded boss	-0.2	0.43	0.67

The pRRSA_c ϕ-parameter optimization results (Table 4) were compared to the median indirectness values across states from experiment 2 (see supplementary material table B.4). This showed, that the evaluations by participants were overall higher (more direct) than the optimized ϕ values.

Model predictions

The parameter values used for the model predictions are shown in table 4. Overall, both models had very similar prediction results and differed only slightly in their predicted distributions. This can be observed for the means across states (columns 4 and 5 in table 3, and fig. 1) as well as in the prediction distributions (see supplementary material). There are some overall differences in the predicted distributions to the experimental results, even though the means over all states are similar as shown in Figure 1. The value $R^2 = 0.595$ was virtually identical for the predictions of both models. Additionally, two different accuracy measures were calculated, as not only the highest predicted outcome (top1-accuracy = 0.634) was of interest, but also the distribution of the second most probable prediction (top2-accuracy = 0.842). Precision and recall values can be found in the supplementary material (Table D.1).

Further, accuracy and R^2 for each relationship can be found in table 5. Here it can be observed, that the model achieved significantly worse results for the *dreaded boss* condition. Overall,

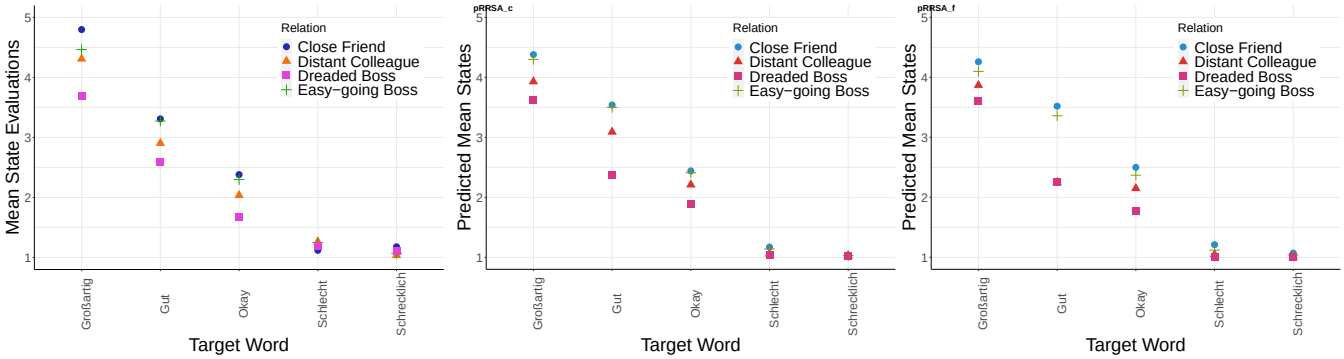

Figure 1: Graphical comparison of behavioral means (experiment 3, left) and model prediction means (pRRSA$_c$, center; pRRSA$_f$ right) over all states for each relationship and target word.

when evaluating the R^2 values, one should consider that it only takes into account the first most probable prediction and not the distribution of the predictions, which is however relevant for the evaluation and can be best observed in distribution graphs (see supplementary material figures D.1).

Model discussion

Both model versions made similar predictions regarding the correct state for the target words across relationships. We expected differences in the prediction performance to identify the parameter that is more suited for the relationship mapping. The choice of parameter for the influence mapping seems, however, to be less relevant than expected. An analysis on which stage of language production – conceptualisation (pRRSA$_c$) or formulation (pRRSA$_f$) – the speaker-influence occurs was thus not possible with this approach. Overall, on average the models predicted similar states compared to the empirical data. However, probability distributions of the predictions differed from the probabilities for the states evaluated by participants.

Some of the results could be explained with the data, as, e.g., for the less accurate predictions for the *dreaded boss* condition, where the data was more distributed, which could be explained with the design of the experiment. An important factor are additional influences not specified in the experimental items (e.g., how long did it take to bake the cake) which might lead to multiple plausible answers in experiments 2 and 3. The variability in interpretation – always present in interaction – as well as listeners' uncertainty when interpreting speech acts (Holtgraves, 2021) likely had an influence as well. In probabilistic models, these should ideally be represented using Bayesian statistics. As the present approach does not allow this, future research should aim at modeling this variability within the relationship influences, e.g., by modeling power and distance explicitly instead of just mapping a general relationship influences to a single model variable (ϕ or α).

Further, we compared the indirectness evaluations from experiment 2 with the optimized pRRSA$_c$ model's ϕ values for each relationship. The differences between these values indicate that the optimized ϕ values might include more (or something different) than just the weighing of directness.

Overall, our modeling approach contributes to formal modeling of politeness, not only by instantiating the RSA-based approach (Yoon et al., 2016) with German data (H-4), but also by adapting it to consider social influences on indirectness.

Conclusions

The research presented in this paper had two goals. The first goal was to find further empirical evidence for the influence of speaker-hearer relations, in terms of power and social distance to the listener, on directness in politeness strategies. The second goal was to instantiate Yoon et al. (2016)'s RSA-based model of politeness with German language data and to extend it by mapping speaker-hearer relation influences to the weighing parameters. Two models were created to compare for which parameter the relationship influence mapping worked better.

The empirical part of our studies found evidence for the influence of speaker-hearer relations on the choice of off-record politeness, confirming previous findings (Holtgraves and Bonnefon, 2017). Our study found influences of power and distance on politeness – previously found by Holtgraves and Yang (1990) for requests – for judgements of the listener's performance in a hobby context. By testing relationships on different ends of the power and distance spectrum, we further found that both factors influence indirect politeness. This is notable because finding strong influences of distance on politeness has proven difficult in previous research (Holtgraves and Bonnefon, 2017; Vergis and Terkourafi, 2015). Future research should analyze this further by also considering different types of distance (frequency of interaction, likeability).

Both pRRSA models – mapping speaker-hearer relations to two different variables in Yoon et al. (2016)'s approach – made similar predictions. It is therefore not possible, with this approach, to state whether relationships should rather influence the speaker strategy in the conceptualisation stage (pRRSA$_c$) or the degree of politeness in the formulation stage (pRRSA$_f$). Overall, both models achieved average predictions similar to the behavioral data, even though the probability distributions over the possible states diverged from the data.

More detailed results and materials are available in the supplementary material: http://doi.org/10.17605/OSF.IO/P4F8C

References

Ambady, N., Koo, J., Lee, F., & Rosenthal, R. (1996). More than words: Linguistic and nonlinguistic politeness in two cultures. *Journal of Personality and Social Psychology*, 70:996–1011. doi: 10.1037/0022-3514.70.5.996.

Brown, P. (2015). Politeness and language. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 326–330. Elsevier, 2nd edition. doi: 10.1016/B978-0-08-097086-8.53072-4.

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

Carli, L. (1999). Gender, interpersonal power, and social influence. *Journal of Social Issues*, 55:81–99. doi: 10.1111/0022-4537.00106.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259, Sofia, Bulgaria.

Frank, M. C. & Goodman, N. D. (2014). Inferring word meaning by assuming the speakers are informative. *Cognitive Psychology*, 75:80–96. doi: 10.1016/j.cogpsych.2014.08.002.

Franke, M. & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44. doi: 10.1515/zfs-2016-0002.

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics*, 14:219–236. doi: 10.1016/0378-2166(90)90081-N.

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interactions. *Psychiatry*, 18:213–231. doi: 10.1080/00332747.1955.11023008.

Goodman, N. D. & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20:818–829. doi: 10.1016/j.tics.2016.08.005.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. L., editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, NY, USA.

Gupta, S., Walker, M. A., & Romano, D. M. (2007). How rude are you?: Evaluating politeness and affect in interaction. In *Proceedings of the 2007 International Conference on Affective Computing and Intelligent Interaction*, pages 203–2017, Berlin, Heidelberg. doi: 10.1007/978-3-540-74889-2_19.

Holtgraves, T. (2021). Understanding miscommunication: Speech act recognition in digital contexts. *Cognitive Science*, 45. doi: 10.1111/cogs.13023.

Holtgraves, T. & Bonnefon, J.-F. (2017). Experimental approaches to linguistic (im)politeness. In *The Palgrave Handbook of Linguistic (Im)politeness*, pages 381–401. Palgrave Macmillan, London, UK. doi: 10.1057/978-1-137-37508-7_15.

Holtgraves, T. & Yang, J.-N. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of Personality and Social Psychology*, 59:719–729. doi: 10.1037/0022-3514.59.4.719.

Kasper, G. (1990). Linguistic politeness: Current research issues. *Journal of Pragmatics*, 14:193–218. doi: 10.1016/0378-2166(90)90080-W.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis. A Tutorial with R, JAGS, and Stan*. Academic Press, London, UK, 2nd edition.

Leech, G. (2014). *The Pragmatics of Politeness*. Oxford University Press, Oxford, UK. doi: 10.1093/acprof:oso/9780195341386.001.0001.

Leichty, G. & Applegate, J. L. (1991). Social-cognitive and situational influences on the use of face-saving persuasive strategies. *Human Communication Research*, 17:451–484. doi: 10.1111/j.1468-2958.1991.tb00240.x.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. The MIT Press, Cambridge, MA, USA.

Scontras, G., Tessler, M. H., & Franke, M. (2021). A practical introduction to the Rational Speech Act modeling framework. *arXiv preprint*, arXiv:2105.09867.

Vergis, N. & Terkourafi, M. (2015). The role of the speaker's emotional state in im/politeness assessments. *Journal of Language and Social Psychology*, 34:316–342. doi: 0.1177/0261927X14556817.

Watts, R. J. (2003). *Politeness*. Cambridge University Press, Cambridge, UK. doi: 10.1017/CBO9780511615184.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 2771–2776, Philadelphia, PA, USA.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the 39th Annual Mee6ng of the Cognitive Science Society*, pages 3602–3607, London, UK.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, 4:71–87. doi: 10.1162/opmi_a_00035.