How to handle the truth: A model of politeness as strategic truth-stretching

Fausto Carcassi (fausto.carcassi@gmail.com)
Department of Linguistics, Keplerstr. 2
72074 Tübingen, Deutschland

Michael Franke (mchfranke@gmail.com)
Department of Linguistics, Keplerstr. 2
72074 Tübingen, Deutschland

Abstract

While the literature has mostly focused on the goal of information transfer, many linguistic phenomena only make sense in the light of further goals pursued by the agent. One such phenomenon is polite language use. In this paper, we propose a new model of polite language production. We suggest that patterns characteristic of polite language, e.g., indirectness, emerge from a tension between two goals: on the one hand, being sufficiently truthful and informative, and on the other hand, being kind to the listener. To capture these pressures, we introduce a novel model of probabilistic language production which combines a strategic choice of content selection with the usual pragmatic choice of content expression. We fit our model to empirical data from a previous experiment using a bespoke Bayesian model. We quantitatively compare our model to a previous model of politeness and discuss some ways in which our account is simpler, more general and better accounts for empirical data and theoretical considerations.

Keywords: politeness; pragmatics; probabilistic modeling; rational speech act; Bayesian data analysis.

Introduction

> Politeness, n: The most acceptable hypocrisy.
>
> Ambrose Bierce, The Devil's Dictionary

In social interactions, we stretch the truth for a variety of reasons, e.g., avoiding punishment, keeping peace of mind, or smoothing out social conflict. In many cases, being liberal with the truth is a rational choice dictated by strategic considerations. Since language is our main tool for communicating (or distorting) the truth, these considerations usually revolve around what to say and how to say it. In this paper, we focus on a specific instance where truth-stretching, social norms, and language production interact: linguistic politeness. We develop an account of when and how a rational speaker with the goal of being polite will resort to stretching the truth. We then implement this mechanism in a Rational Speech Act (RSA) model of pragmatic communication (Frank & Goodman, 2012; Goodman & Frank, 2016; Franke & Jäger, 2016), and show that it can account for experimental data from previous work in politeness.

The most developed cognitive model of polite language to date is by Yoon, Tessler, Goodman, and Frank (2020), who build on much previous work on understanding politeness within the framework of Gricean pragmatics (See e.g., Lakoff (1973); Brown and Levinson (1978); Leech (1983) for some classic references on politeness). They argue that the roots of linguistic politeness can be traced to goal-directed pragmatic reasoning (see also work by van Rooij (2003) or Mühlenbernd, Zywiczynski, and Wacewicz (2019) for like-minded approaches). Specifically, polite utterances are an attempt at finding the best compromise between three possibly competing communicative goals. The first goal in polite language production is to be informative - this Yoon et al. (2020) dub *informational* goal, consistently with classic Gricean pragmatics. A second goal for a polite speaker is to make the listener feel good; Yoon et al. (2020) call this goal the *prosocial* goal. Ideally, informational and prosocial utilities would always be aligned. However, sometimes they conflict with each other, because an informative and truthful speaker would have to say something that makes the listener feel bad. In such cases, the speaker aims to present themselves as someone who cares about both prosocial and informational utility, but cannot maximize them both at the same time; Yoon et al. (2020) call this the *self-presentational goal*. Yoon et al. (2020) argue that finding a balance between these three goals is the cause of indirect polite speech.

We briefly illustrate this mechanism with an example. Suppose a friend of yours has made a cake and you tried a slice of this cake, and now your friend is asking what you thought of the cake. In an ideal situation you would be able to truthfully say that you found the cake amazing. However, you in fact thought that the cake was quite poor. So you are confronted with a choice: you can either say something that does not reflect your experience, e.g., "the cake was good", or express your true feelings about the cake but risk hurting your friend, e.g., "the cake was very bad". To get out of this conflict, you choose to say something that signals to your friend that you want to be both kind and truthful but cannot quite do both at once: "The cake is not terrible". This last utterance is truthful, but underinformative.

The self-presentational model of politeness is a model of *how* to convey some content. However, this is not the whole story: politeness involves both choosing the right words to describe a situation, and deciding what to describe in the first place. Consider again the cake scenario above. Since your friend has no way of knowing how much you truly liked the cake, instead of signaling that you are attempting to be kind, you can also decide to tell a white lie by exaggerating your appreciation. How much should you pretend to like the cake? You have to choose a level of pretense that strikes a balance

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

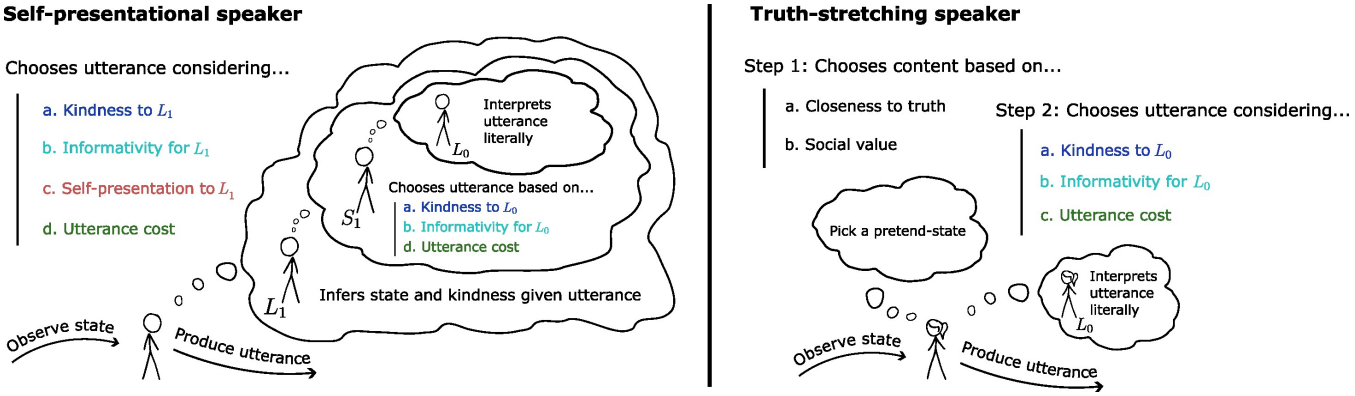


Figure 1: Schematic comparison between the speakers in the self-presentational (left half of the plot) and in the truth-stretching (right half of the plot) models. Both speakers observe a true state and, based on this observation, produce an utterance. However, the factors involved in the utterance choice is different in the two models.

between kindness and realism: Your friend would likely not believe too positive an opinion, but if you accurately describe the cake they will be hurt. So you consider the trade-off between kindness and accuracy for each possible pretend-state (hypothetical degree of cake appreciation). Slightly more technically, the probabilistic model of content choice we introduce below predicts that you will sample a pretend-state to describe from a probability distribution, where states that balance social value and realism are more likely.

Once you selected a pretend-state, you can then choose the kindest utterance conditioned on its pretended truth. For instance, if you pretend that the cake is okay, it is kinder to say "The cake is not bad" than to say "The cake is okay". Arguably, this is because the states compatible with the former have a higher expected social value than for the latter.

More in general, in order to maximise the social value of their utterances speakers may intentionally assert falsehoods. Not brutal lies necessarily, maybe just a little false. This mechanism, which we call *truth-stretching*, allows for kindness in the face of an undesirable reality. The truth-stretching model of politeness introduced in this paper therefore features a double choice of the speaker: at the level of content, and at the level of content expression. This is a fairly general, and arguably independently useful, extension to recent probabilistic models of pragmatic language use which have so far almost exclusively focused on the latter. Figure 1 shows a schematic comparison between the two accounts.

Various clarifications are in order. First, there are many factors that modulate how much truth can be stretched. In the example above, the listener's uncertainty about the speaker's taste allows for greater truth-stretching: since there is variation in people's taste, the listener might happen to be one of the people who do not mind the ways in which the speaker's cake is bad. Power dynamics and social status can also play a role; a more powerful person for instance might know that they likely will not be called out for saying something implausible (Gu et al., 2020). Truth-stretching need not even be misleading. Participants in a conversation are sometimes expected to cooperate in a lie, as expressed by the Russian term *vranyo*: "A Russian friend explained *vranyo* this way: 'You know I'm lying, and I know that you know, and you know that I know that you know, but I go ahead with a straight face, and you nod seriously and take notes.'" (Shipler, 2016). On the other hand, lying is effort (Verschuere, Köbis, Bereby-Meyer, Rand, & Shalvi, 2018), and lying more is more effort. Since the factors that influence the extent to which truth can be stretched are open-ended, we will not encode them directly in the model, and leave a formal description of the main motivations to future work.

Second, the truth-stretching example illustrates that social value—in this case, how happy your friend would be for each level of appreciation—plays a *double duty* in determining polite utterances. First, in content choice: everything else being equal, you are more likely to pretend to be in a more socially desirable state. Second, in content expression: given a pretend-state, it is quite likely that you will pick an utterance that includes at least some desirable states.

Third, the truth-stretching account presents some advantage over a self-presentational account. First, self-presentation does not play a role for truth-stretching. This is important because self-presentational utility requires two levels of nested recursive mindreading, while truth-stretching only requires one. Empirical work tells us that polite truth-stretching arguably appears before deeply nested recursive mindreading: children as young as 5 tell white lies in order to be kind (Talwar, Murphy, & Lee, 2007), and often even approve of lying for social reasons (Walper & Valtin, 1992). Second, our picture is intuitively more consistent with the rest of our social behaviour. Suppose that a friend asks you if you like their present to you, and you do not like the present, so you make a scene of pretending to wipe sweat off your forehead. This behaviour clearly signals that (1) you did not like the present, (2) you do not want to say that you did not like the present, and (3) you care about being truthful while also

not insulting your friend's present. However, this is strange *as polite behaviour*; your friend might (rightly) wonder why you did not simply pretend to like the present.

The self-presentational model of politeness

The self-presentational model of politeness by Yoon et al. (2020) is specified as a Rational Speech Act (RSA) model. RSA is a framework where pragmatic language production is modelled in agents capable of utility maximization and probabilistic inference. A typical RSA model of language production starts with a literal listener L_0, who computes a distribution over world states s by conditionalizing on the truth of a received utterance u (assuming uniform priors over s here):

$$P_{L_0}(s \mid u) \propto \delta_{s \in [\![u]\!]} \tag{1}$$

The first-order pragmatic speaker S_1 observes the world state s and calculates the utility of each utterance, which is higher when the utterance is more informative to the literal listener, and lower when it requires more effort to produce:

$$U_{S_1}(u;s) = \log P_{L_0}(s \mid u) - C(u) \tag{2}$$

S_1 then tends to choose utterances with high utility given the observed world state, as the result of a softmax operation, modulated by an α parameter:

$$P_{S_1}(u \mid s) \propto \exp\left[\alpha U_{S_1}(u;s)\right] \tag{3}$$

When $\alpha = 0$, the distribution is uniform. As $\alpha \to \infty$, it becomes more peaked on the signals with the highest utility.

Finally, the pragmatic listener does Bayesian update to calculate the probability of each possible state given the utterance they received and the probability that the pragmatic speaker would have produced that utterance in each state:

$$P_{L_1}(s \mid u) \propto P_{S_1}(u \mid s) \tag{4}$$

Higher levels of recursive mindreading can be easily obtained by alternating speakers at level t who optimize the utility of each signal given the observed state considering a listener at level $t-1$, and listeners at level $t+1$ who run Bayesian inference over states given a message with a likelihood function encoded by a speaker at level t.

The self-presentational model of politeness extends this model by assuming a graded semantics, where the compatibility of utterance u with state s, $\mathcal{L}(u,s) \in [0;1]$, ranges from 0 (completely incompatible) to 1 (completely compatible):

$$P_{L_0}(s \mid w) \propto \mathcal{L}(u,s) \tag{5}$$

The first-order pragmatic speaker S_1 then calculates the utility of a signal by striking a balance between three utility components. First, the informativity of an utterance for the literal listener, in a way similar to the RSA speaker S_1 above. Second, the social utility of the utterance, which is defined as the expected desirability of the true state given the utterance. Third, the utterance's production cost. Production probabilities are defined as above as a function of the utility:

$$P_{S_1}(w \mid s, \phi) \propto \exp[\alpha U_{S_1}(w;s)] \tag{6}$$

$$\begin{aligned} U_{S_1}(w;s) =& \phi \ln P_{L_0}(s \mid w) \\ &+ (1-\phi)\mathbb{E}_{P_{L_0}(s' \mid w)}[V(s')] - C(w) \end{aligned} \tag{7}$$

The first-order pragmatic listener L_1 uses uniform priors over ϕ to obtain the joint distribution of s and ϕ from Bayes rule, given utterance u produced by the pragmatic speaker S_1:

$$P_{L_1}(s, \phi \mid w) \propto P_{S_1}(w \mid s, \phi) \tag{8}$$

The second-order pragmatic speaker S_2 is similar to the first order S_1, with two crucial differences. First, S_2 calculates the social value and informativeness using the distribution over states given the utterance for L_1 rather than L_0. Second, S_2 considers the presentational utility, which is computed as the probability, given the utterance, of S_2's real level of politeness from the perspective of the first-order listener L_1. Formally:

$$\begin{aligned} U_{S_2}(w;s;\omega;\phi) =& \omega_{\text{inf}} \ln P_{L_1}(s \mid w) + \omega_{\text{soc}} \mathbb{E}_{P_{L_1}(s' \mid w)}[V(s')] \\ &+ \omega_{\text{pres}} P_{L_1}(\phi \mid w) - C(w) \end{aligned} \tag{9}$$

$$P_{S_2}(w \mid s, \omega, \phi) \propto \exp[\alpha U_{S_2}(w;s;\omega;\phi)] \tag{10}$$

Where $P_{L_1}(s \mid w) = \int_0^1 P_{L_1}(s, \phi \mid w) d\phi$ is the probability of the rational listener guessing state s given utterance u, across all values of ϕ.

Yoon et al. (2020) experimentally tested the predictions of their model. A first experiment estimated the graded semantics encoded by $\mathcal{L}$ by asking participants whether each combination of state and utterance was compatible. In the main experiment, 202 participants were presented with a scenario where they had to give feedback concerning some performance or product, e.g., a cake. In each trial, the participants were given a true level of appreciation between zero and three hearts. The feedback could be any one of eight utterances: "terrible", "bad", "good", "amazing", and their negations. Each participant saw four different scenarios (one for each level of appreciation) for each of three conditions, for a total of twelve scenarios per participant. The three conditions manipulated different speaker goals: to be *informative* ("Give accurate and informative feedback"), to be *kind* ("Make the listener feel good"), and to be *both* informative and kind.

We implemented the model in PyMC (Salvatier, Wiecki, & Fonnesbeck, 2016) and fitted it to the data.[1] Our results closely mirror those reported by Yoon et al. (2020). Posterior predictive distributions are shown in Figure 2 along with the data. The crucial empirical hypothesis of Yoon et al. (2020) was that participants in the *both* condition would naturally consider the self-presentational goal in their utterance choice.

[1] All the code for the models described in this paper is available at `https://github.com/thelogicalgrammar/politeness`.

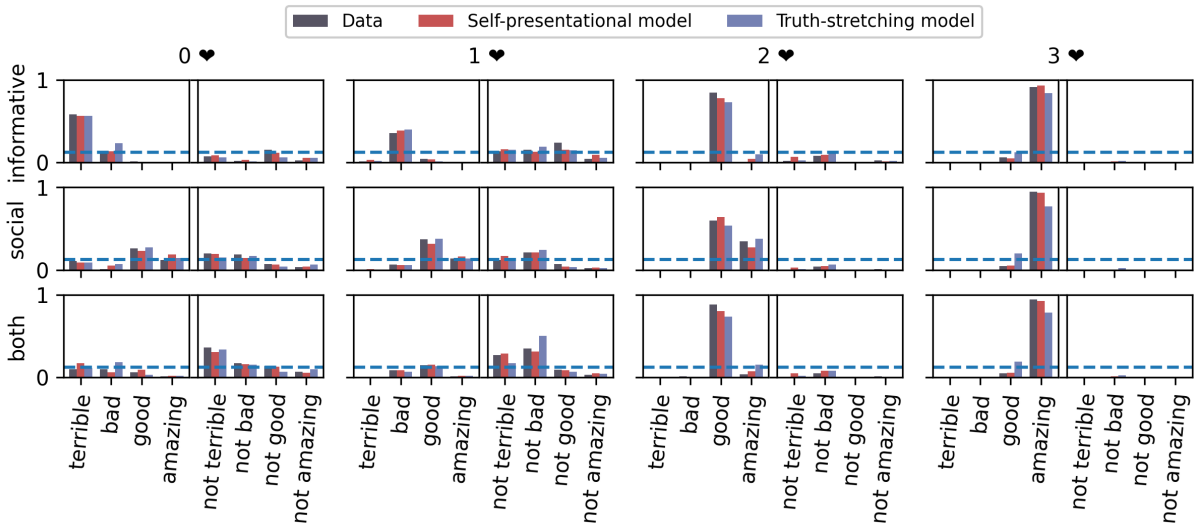


Figure 2: Comparison of the experimental results with the posterior predictive samples of the self-presentational model and the truth-stretching model, aggregated across participants. Each group of plots shows a state. The left columns of plots show positive utterances and the right columns negated utterances. For the models, the plot shows the mean production probabilities of polite speaker $S_2(w \mid s)$ in the posterior predictive samples, for the eight utterances and four states in the experiment. Both models capture the general patterns in aggregated production probabilities seen in the data.

For zero and one hearts, participants in the *both* condition used negative utterances (such as "not terrible"), and when fitting $\mathcal{L}$, C, α, ϕ, and ω, the model is able to closely account for the data. In the interest of space, we point to the original paper for a fuller discussion of these results.

A computational model of truth-stretching

The main conceptual innovation of the truth-stretching account is a model of the speaker's utterance-generation process that proceeds incrementally: first, for the true state s a "pretend state" s' is sampled with probability $P(s' \mid s)$, and then, with probability, $P_{S_1}(u \mid s')$, the utterance is selected for the "pretend state" s.[2] The resulting overall probability of utterance u for state s is just the sum of all path-probabilities via all intermediate "pretend states" s':

$$P_{S_1}(u \mid s) = \sum_{s' \in S} \left(P_{S_1}(u \mid s')\, P(s' \mid s) \right) \tag{11}$$

Previous work has used exactly this approach of two-stage utterance generation in the context of vagueness (Franke & Correia, 2018; van Tiel, Franke, & Sauerland, 2021, 2022), where the transition probabilities $P(s' \mid s)$ are due to noise (e.g., imprecise perception). The main innovation of the truth-stretching model is to consider this a separable strategic choice of the speaker at the level of content selection similar to classic approaches in natural language generation (Gatt & Krahmer, 2018).

For a truth-stretching model of polite language use, we assume that strategic content selection, captured by probabilities $P(s' \mid s)$, is sensitive to the distance between s' and s in semantic space (Carcassi & Szymanik, 2021); states that are more distant from the observed state are less likely to be selected. We model distance as the negative quadratic difference between the true state s and the chosen content s'. Moreover, the speaker wants to pretend to be in a state with high social utility, encoded in the value function V. In the case example above V might for instance encode how much the speaker liked the cake. The resulting definition is, where φ modulates the trade-off between value and distance, and θ is a softmax parameter:

$$P(s' \mid s) \propto \exp\left[\theta\left((1-\varphi)V(s') - \varphi(s'-s)^2\right)\right] \tag{12}$$

Utterance choice for a "pretend state," $P_{S_1'}(u \mid s')$, is defined, similar to the above, based on a mixture of three utility components, i.e., politeness, informativity, and simplicity (where ω is a probability vector and L_0 is defined as in Equation 1):

$$P_{S_1'}(u \mid s) \propto \exp\left[\alpha U_{S_1}(u;s)\right] \tag{13}$$

$$U_{S_1}(u;s) = \omega_1 \mathbb{E}_{P_{L_0}(s' \mid u)}\left[V(s')\right] + \omega_2 \log P_{L_0}(s \mid u) - \omega_3 C(u) \tag{14}$$

Figure 3 shows the production probabilities of the polite truth-stretching speaker for one of the scenarios from the experiment described above, for a specific setting of the parameters. Various patterns can be seen in the Figure. When the

[2] It is also conceivable to formulate a model in which the speaker chooses both u and s' simultaneously, so that the choice of "what to say" is also conditioned by how efficiently a message can be communicated. Franke and Bergen (2020) explore such a joint-choice model for a related but different case, namely where the speaker jointly chooses an utterance and an "intended meaning" of that utterance. We leave a comparison of such variants for future work.

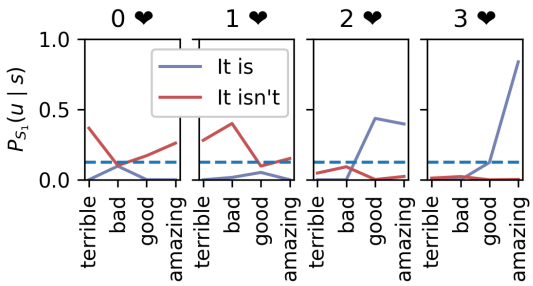


Figure 3: Each of the four subplots shows the production probabilities of the polite truth-stretching speaker conditioned on a true like-state (zero to four states) for the positive (blue) and negated (red) utterances. Production probabilities are marginalized across pretend-states, as described in Equation 11. Parameters are set as follows: $V = [-10, -5, 0, 3, 5]$, $\alpha = 2$, $\varphi = 0.7$, $\theta = 2.$, $\omega \propto [1, 1, 1]$, and were chosen by hand for illustrative purposes.

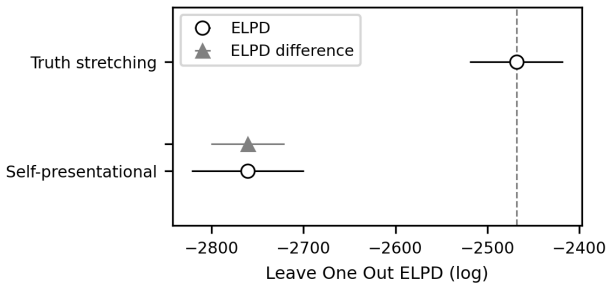


Figure 4: Quantitative comparison of the two models according to an approximation of their predictive accuracy. The hierarchical truth-stretching model is better than the pooled self-presentational model with respect to ELPD (Expected Log Pointwise Predictive Density, which estimates out-of-sample predictive accuracy) of the data in the main experiment.

speaker genuinely likes the cake, they are very likely to say the cake is "amazing", and when they consider the cake good but not excellent, they sometimes say "good" and sometimes they stretch the truth and say "excellent" to be kind. The extent to which the speaker is ready to stray from the truth depends on how much is gained in terms of social value, and the social returns are likely to diminish for more positive values: the trade-off may be worth to say "not bad" instead of "terrible", but not to say "excellent" instead of "good".

In both of the positive states in Figure 3, the speaker almost exclusively uses positive utterances. On the other hand, when the speaker considers the cake terrible, they almost never say "terrible". Rather, they resort to three main strategies: (1) using an indirect utterance that is compatible with the truth, such as "not good" and "not amazing", (2) stretching the truth and minimize their dislike, by saying it is merely "bad", (3) combining the first two strategies and produce "not terrible". Overall, the model accounts for the intuitive fact that negative states lead a polite speaker to produce more negated, indirect utterances.

Data fitting and model comparison

In order to compare the self-presentational and the truth-stretching model, we also fitted the latter to the data from the production experiment in Yoon et al. (2020). We do not fit a graded semantics for the signals, but rather assume that the speaker simply draws an intuitive threshold semantics for the four utterances: $[\![\text{terrible}]\!] = \{0\}$, $[\![\text{bad}]\!] = \{0, 1\}$, $[\![\text{good}]\!] = \{2, 3\}$, $[\![\text{amazing}]\!] = \{3\}$, and implement negation as difference from the set of all states. Moreover, we assume that the speaker imagines an additional, neutral state between "bad" and "good", for which the experiment lacks a signal.[3]

[3]Without such a neutral state, "not bad" and "good' become synonymous. It is easy to imagine that participants imagine a neutral state between "bad" and "good" if they think the receiver of their message may not have access to the particular set of states and signals in the experiment. This is particularly plausible for this experimental design, since like-states are meant to be the participant's subjective evaluations.

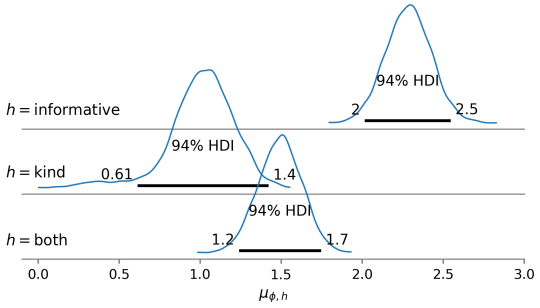
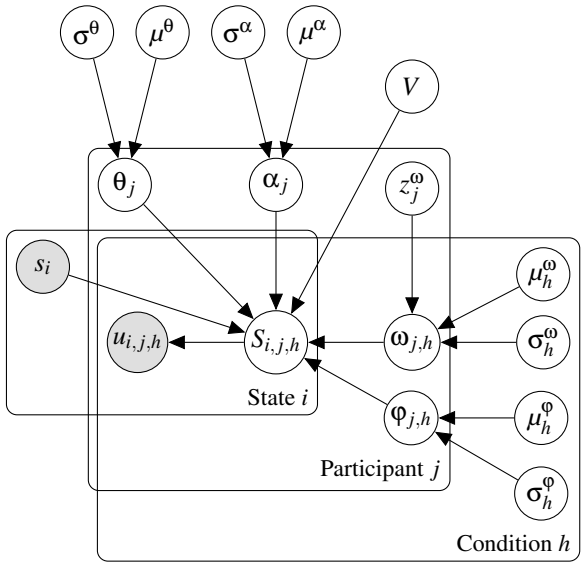


Figure 5: Posterior over the population-level value of μ_h^φ for the three conditions h (see Figure 6 for a fuller description of the meaning of the parameter). Parameter μ_h^φ encodes, for each of the three conditions h, a population-level baseline for the tendency to stick close to the observed state. The higher the value of μ_h^φ, the stronger the tendency to select the true state or pretend-states close to it.

We define the model with the hierarchical structure described in Figure 6. We ran one thousand tuning steps and one thousand samples in four chains. Figure 2 shows the posterior predictive distribution of the truth-stretching model, aggregated across participants. The aggregate predictions are close to the observed data, showing that the model can capture the patterns discussed above. Moreover, the hierarchical structure allows our model to capture individual variation in production behaviour, which turns out to be important for overall fit. Figure 4 shows a quantitative comparison between the two models in terms of an approximation to their predictive accuracy on out-of-sample data from the production



(a) DAG for model

$$\mu^\alpha \sim \mathcal{N}(0,1) \quad \sigma^\alpha \sim \mathcal{N}^+(1)$$
$$\log(\alpha_j) \sim \mathcal{N}(\mu^\alpha, \sigma^\alpha) \quad \text{for } j \in 1,...,202$$
$$\mu_h^\omega \sim \mathcal{N}^+(\mathbf{1}_3, \mathbb{1}_{3,3}) \quad \sigma_h^\omega \sim \mathcal{N}^+(\mathbf{1}_3, \mathbb{1}_{3,3}) \quad \text{for } h \in 1,2,3$$
$$z_j^\omega \sim \mathcal{N}^+(0,1) \quad \text{for } j \in 1,...,202$$
$$\omega_{j,h} \sim \text{Dir}(\mu_h^\omega + \sigma_h^\omega z_j^\omega)$$
$$\mu_h^\varphi \sim \mathcal{N}(0,1) \quad \sigma_h^\varphi \sim \mathcal{N}^+(1) \quad \text{for } h \in 1,2,3$$
$$\text{expit}(\varphi_{j,h}) \sim \mathcal{N}(\mu_h^\varphi, \sigma_h^\varphi) \quad \text{for } j \in 1,...,202$$
$$\mu^\theta \sim \mathcal{N}(0,1) \quad \sigma^\theta \sim \mathcal{N}^+(1)$$
$$\log(\theta_j) \sim \mathcal{N}(\mu^\theta, \sigma^\theta) \quad \text{for } j \in 1,...,202$$
$$V \sim \text{Ord}\,\mathcal{N}(\mathbf{1}_5, 5\,\mathbb{1}_{5,5})$$
$$S_{i,j,h} = P_{S_1}(\cdot \mid \alpha_j, \theta_j, V, \omega_{j,h}, \varphi_{j,h}, s_i)$$
$$u_{i,j,h} \sim \text{Cat}(S_{i,j,h})$$

(b) Model description

Figure 6: Description of the truth-stretching model for analyzing the data from Yoon et al. (2020). $u_{i,j,h}$ is the utterance produced by the participant j in condition h after observing state s_i. $S_{i,j,h}$ is the probability vector encoding the production probabilities for participant j, condition h, and state i, as per Equation 11 spelled out here with explicit parameterization. $\mathcal{N}$ is a normal distribution, $\mathcal{N}^+$ is a half-normal distribution, Ord $\mathcal{N}$ is a normal distribution with ordered constraint, Dir is a Dirichlet distribution, and Cat is a categorical distribution taking as parameter an (unnormalized) vector. Function expit is the inverse function of logit, where $\text{logit}(p) = \log(p/(1-p))$.

experiment (Vehtari, Gelman, & Gabry, 2017).[4]

The effect of the different goals in the three conditions is reflected in a natural way in the posterior distributions of the fitted parameters. Figure 5 shows the posterior distribution of μ_h^φ, the population-level parameter encoding the resistance against stretching the truth; the *kind* condition allows for strong truth-stretching, the *informative* condition precludes truth-stretching, and the *both* condition asks the agent to find a balance between the two.

Conclusions

Politeness is a multifactorial phenomenon; there is no one reason *for*, and no one way *of* being polite. Previous work has uncovered self-presentation as one of the mechanisms that underlie politeness. In this paper, we have argued that stretching the truth is another and in some ways more fundamental mechanism. What is the relation between them? In essence, self-presentational agents signal to the listener a struggle between balancing social and informational utility, while truth-stretching speakers simply pretend, as much as they feel they can get away with, that the struggle is minimal or does not exist in the first place. The agent in the self-presentational model resolves a tension between being truthful and being kind by showing that they are at an impasse; a truth-stretching agent resolves the tension by telling a white lie. Future work can integrate both mechanism in a single underlying model that predicts which of the two strategies a speaker will use, depending on features of the context of utterance.

More generally, there is added value in the truth-stretching model in that it is, to the best of our knowledge, the first RSA model that incorporates a strategic layer of "content choice" prior to the selection of an utterance to encode a message. This is independently useful and opens applications to other phenomena beyond politeness, such as for strategic uses of vagueness (de Jaegher & van Rooij, 2011), imprecision (Krifka, 2007), or pragmatic slack (Lasersohn, 1999).

Moreover, laterally comparing model variants, even if they target the same phenomena and even data sets, is important for a maturing field like probabilistic pragmatics (Marcus & Davis, 2013). In this particular comparison between the truth-stretching and the self-presentation model, it seems that not only conceptual arguments speak in favor of the former, but also appeal to simplicity. Higher-order recursion is *not* necessary to get similar, if not better results.

[4]The comparison presented above is between a hierarchical version of our model and a completely pooled version of the self-presentational model. Unfortunately, the integration needed to compute $P_{L_1}(s \mid w)$ presents a technical difficulty in implementing a hierarchical version of the latter. While in the current version we approximate the integral via grid approximation, this strategy is not scalable to a hierarchical model. We plan to work on this technical issue in future work.

Acknowledgments

We would like to thank Polina Tsvilodub and Juliane Schwab for helpful discussions. This project has received funding from the European Union's Horizon 2021 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101065866 (PlauDe) awarded to Fausto Carcassi. Michael Franke is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764.

References

Brown, P., & Levinson, S. C. (1978). Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction* (pp. 56–311). Cambridge University Press.

Carcassi, F., & Szymanik, J. (2021). An alternatives account of 'most' and 'more than half'. *Glossa: A Journal of General Linguistics*, *6*(1). doi: 10.16995/glossa.5764

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. doi: 10.1126/science.1218633

Franke, M., & Bergen, L. (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, *96*(2), e77-e96. doi: 10.1353/lan.2020.0034

Franke, M., & Correia, J. (2018). Vagueness and imprecise imitation in signaling games. *The British Journal for the Philosophy of Science*, *69*(4), 1037–1067.

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

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi: 10.1016/j.tics.2016.08.005

Gu, Z., Liu, L., Tan, X., Liang, Y., Dang, J., Wei, C., … Wang, G. (2020). Does power corrupt? The moderating effect of status. *International Journal of Psychology: Journal International De Psychologie*, *55*(4), 499–508. doi: 10.1002/ijop.12629

de Jaegher, K., & van Rooij, R. (2011). Strategic vagueness, and appropriate contexts. In *Language, games, and evolution* (pp. 40–59). Berlin, Heidelberg: Springer.

Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 111–126). Amsterdam: KNAW.

Lakoff, R. (1973). Language and Woman's Place. *Language in Society*, *2*(1), 45–80.

Lasersohn, P. (1999). Pragmatic halos. *Language*, *75*(3), 522–551.

Leech, G. N. (1983). *Principles of pragmatics* (No. 30). London ; New York: Longman.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychol Science*, *24*(12), 2351–2360.

Mühlenbernd, R., Zywiczynski, P., & Wacewicz, S. (2019). *The Game Theory of Politeness in Language: A Formal Model of Polite Requests* (Preprint). PsyArXiv. doi: 10.31234/osf.io/6srux

van Rooij, R. (2003). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In *Tark: Proceedings of the 9th conference on theoretical aspects of rationality and knowledge* (pp. 45–58). New York: ACM.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. doi: 10.7717/peerj-cs.55

Shipler, D. K. (2016). *Russia: Broken Idols, Solemn Dreams (Revised Edition)*. Crown.

Talwar, V., Murphy, S. M., & Lee, K. (2007). White lie-telling in children for politeness purposes. *International Journal of Behavioral Development*, *31*(1), 1–11. doi: 10.1177/0165025406073530

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*.

van Tiel, B., Franke, M., & Sauerland, U. (2022). Meaning and use in the expression of estimative probability. *Open Mind: Discoveries in Cognitive Science*, *6*, 250—263.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018, September). Taxing the Brain to Uncover Lying? Meta-analyzing the Effect of Imposing Cognitive Load on the Reaction-Time Costs of Lying. *Journal of Applied Research in Memory and Cognition*, *7*(3), 462–469. doi: 10.1016/j.jarmac.2018.04.005

Walper, S., & Valtin, R. (1992). Children's understanding of white lies. In R. J. Watts, S. Ide, & K. Ehlich (Eds.), *Politeness in Language* (pp. 231–252). De Gruyter Mouton. doi: 10.1515/9783110886542-012

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