Is the asymmetry in negative strengthening the result of adjectival polarity or face considerations?

Sarang Jeong, Christopher Potts, Judith Degen
{sarangj, cgpotts, jdegen}@stanford.edu
Department of Linguistics, Stanford University
Stanford, CA 94305 USA

Abstract

Sentences with negated adjectives receive a stronger interpretation than given by their semantics, a phenomenon called negative strengthening. It has been reported that inherently positive adjectives display a higher degree of negative strengthening than inherently negative adjectives. We investigate two possible causes of this asymmetry: intrinsic adjectival polarity and face considerations. Results of an experiment where face-related factors were manipulated suggest that both polarity and face contribute to the asymmetry. Extending a probabilistic RSA model of polite speech, we formalize the listener's reasoning about a speaker's use of negated adjectives as a tradeoff between expecting a speaker to maximize both an utterance's social and informational utility, while avoiding inherently costly adjectives.

Keywords: negative strengthening; gradable adjectives; pragmatic inference; Rational Speech Act

Introduction

Sentences with negated adjectives, like "Sam's talk wasn't good" give rise to the phenomenon of *negative strengthening*: rather than simply negating that the talk was good, the speaker is taken to convey that Sam's talk was in fact bad. That is, sentences with negated adjectives receive a stronger interpretation than given by their semantics (Brown & Levinson, 1987; Gotzner & Mazzarella, 2021; Horn, 1989; Mazzarella & Gotzner, 2021; Ruytenbeek et al., 2017). However, inherently negative adjectives have been noted to not give rise to the same amount of negative strengthening as inherently positive adjectives. For example, "Sam's talk wasn't bad" is not taken to mean that Sam's talk was in fact good. The source of this asymmetry in negative strengthening has variably been attributed to the *polarity* of the produced adjectives (Mazzarella & Gotzner, 2021; Ruytenbeek et al., 2017) or to *face considerations* involved in the use of positive vs. negative adjectives (Brown & Levinson, 1987; Gotzner & Mazzarella, 2021; Horn, 1989; Yoon et al., 2020). The present study seeks to shed light on this issue.

Under the *Polarity Hypothesis*, the asymmetry in negative strengthening is attributed to adjectives' intrinsic polarity. The explanation rests on four crucial assumptions: first, that antonym pairs of adjectives differ in intrinsic polarity (e.g., *good* is considered inherently positive and *bad* inherently negative, Cruse, 1986; Horn, 1989; Lehrer & Lehrer, 1982); second, that negative polarity adjectives are inherently more complex than their positive antonyms (Büring, 2007a,b); third, that negated adjectival expressions ("not good", "not bad") are more complex or costly to produce than their non-negated counterparts ("good," "bad") (Carpenter & Just, 1975; Clark & Chase, 1972; Just & Carpenter, 1971, 1976, see Kaup & Dudschig, 2020 for an overview); and fourth, simple/cheap expressions are used to describe clear cases while complex/costly expressions are used to describe atypical cases (division of pragmatic labor, Horn, 1984). Under these assumptions, the listener's reasoning is assumed to go as follows:

A simple expression like "Sam's talk was bad" would mean it was clearly bad while a more complex expression like "Sam's talk wasn't good" would mean it was an atypical case of the antonymic state (*bad*), i.e., only somewhat bad. Then, expressions like "Sam's talk wasn't bad," being even more complex due to the overt negation and the negative polarity of the adjective, would indicate a highly atypical case of the antonymic state (*good*), i.e., far from typically good. Under this reasoning, the inferred meaning of "wasn't good" (*somewhat bad*) is closer to the antonym than the inferred meaning of "wasn't bad" (*far from good*) is, creating the polarity asymmetry in negative strengthening (Krifka, 2007; Ruytenbeek et al., 2017).

Under the *Face Hypothesis*, the asymmetry in negative strengthening is attributed to the listener's reasoning about face considerations the speaker engages in, on the assumption that they are following Grice (1975)'s cooperative principle. Face refers to the self-image of the speaker or listener, and it is assumed that the speaker's goal is not only to convey information but also to save face (Brown & Levinson, 1987; Lakoff, 1973; Leech, 1983; Pfister, 2010). The listener, upon observing "Sam's talk wasn't good," reasons that the speaker could not have said "Sam's talk was bad" if they had meant it, because the described state threatens Sam's face (if the speaker cares about that) or their own (if they do not want to be perceived as maligning a third party). Thus, by using a more costly, less informative, yet less face-threatening alternative ("not good"), the speaker is likely to want to convey the face-threatening state indirectly. This reasoning does not apply to "Sam's talk wasn't bad" because there is no face-based reason to avoid its less costly, more informative alternative ("good"), so the most informative resulting interpretation is that Sam's talk was neither good nor bad (Mazzarella & Gotzner, 2021).

While a number of studies report a polarity asymmetry in negative strengthening (Colston, 1999; Fraenkel & Schul, 2008; Gotzner & Mazzarella, 2021; Mazzarella & Gotzner, 2021; Ruytenbeek et al., 2017; Yoon et al., 2017, 2020),there is little work that directly compares adjectival polarity and face considerations as the driving forces behind the asymmetry. One exception is Mazzarella & Gotzner (2021). Mazzarella & Gotzner pitted face and polarity against each other by attempting to reverse the value scale, or desirability of the described state. In 'ordinary' contexts, negative adjectives conveyed face-threatening states; in 'non-ordinary' contexts, they conveyed face-*saving* states. The results suggested that there was greater negative strengthening for positive than negative adjectives regardless of context, which the authors interpreted as supporting the Polarity Hypothesis.

However, several design choices limit the conclusiveness of that study. First, the stimuli did not contain non-negated forms ("good," "bad"), not allowing a direct inference regarding how much the negated forms were *strengthened* in relation to the non-negated forms. More importantly, there was no independent manipulation check – i.e., no independent evidence that participants really perceived the value scale to flip in ordinary and non-ordinary contexts. This is concerning given that, based on researcher intuition, the nature of the intended value scales in the non-ordinary contexts was often ambiguous. For instance, the following is a non-ordinary context used by Mazzarella & Gotzner: *You want to go to the casino with your friend even though you have a gambling addiction. That day the casino is closed. Your friend tells you: "You are not lucky/unlucky"*. While intended to make the unlucky state the desirable one, it is not clear that that manipulation succeeded.

Our contribution in this paper is two-fold: first, inspired by Mazzarella & Gotzner, we conduct a behavioral experiment pitting the Polarity and Face Hypothesis against each other, with a clearer manipulation of the value scale. Second, we model the behavioral results by modifying the polite Rational Speech Act (RSA) model of Yoon et al. (2016, 2017, 2020), which formalizes polite language by positing a speaker with both informational and social goals. Specifically, we allow the value function to vary depending on the context, and introduce an adjectival polarity cost. The model reproduces the major qualitative empirical patterns by allowing for *both* adjectival polarity and face considerations to drive the asymmetry in negative strengthening.

Experiment[1]

To examine whether polarity or face considerations cause the asymmetry in negative strengthening, we manipulated both the observed adjectives (differing in polarity) and the contextually given value scale.[2]

Both the Polarity and the Face Hypothesis predict an asymmetry in negative strengthening by adjectival polarity. However, they differ in the explanation of the prediction. Under the Polarity Hypothesis, adjectival polarity directly drives the asymmetry. Thus, reversing the value scale should not lead to a change in the asymmetry (reflecting the results of Mazzarella & Gotzner, 2021). In contrast, under the Face Hypothesis, it is face considerations that drive the asymmetry. Thus, reversing the value scale should also reverse the asymmetry: there should be more negative strengthening for negative than for positive polarity adjectives when the value scale is reversed.

Methods

Participants We recruited 240 participants through Prolific. After excluding those who self-reported a non-English native language (N=5) and those who gave more than 2 incorrect answers to the VALUE QUESTION in a main stimulus or the STATE QUESTION in a control stimulus (details below) (N=4), 231 participants remained.

Procedure and materials A main trial consisted of 3 components, shown in Figure 1.A. In the first component, participants read a brief context that set up the value scale, including via an explicit statement (***speaker** wants **entity** to be **adj***), where **speaker** was a contextually established speaker, **entity** was either a human or non-human entity[3] to be described in a target utterance, and **adj** was drawn from four antonym pairs (*good/bad, big/small, long/short, fast/slow*). They then answered the VALUE QUESTION that functioned as a manipulation check assessing the perceived value scale (*What does **speaker** want **entity** to be like?*). Participants provided responses on a sliding scale with endpoints labeled "very **neg-adj**" to "very **pos-adj**," where **pos-adj** and **neg-adj** formed an antonymic pair (e.g., *fast/slow*).

In the second component, participants read a short sentence establishing a collegial relation between the speaker and their interlocutor, followed by the speaker's target utterance: "**entity** was/wasn't **adj**."[4] They then answered the STATE QUESTION that probed their interpretation of the sentence (*What was **entity** like?*) by providing responses on a sliding scale with endpoints labeled "very **neg-adj**" to "very **pos-adj**."

Finally, the third component consisted of two GOAL QUESTIONS (HONESTY and POSITIVITY), each of which measured the speaker's informational and social goals as inferred by

[1] Methods, exclusions, and analyses for the experiment were preregistered through the Open Science Foundation at `https://osf.io/z4s6f`. The experiment, data, analyses, and figures reported in this study are available at `https://github.com/sarangjeong/negated-adjectives`.

[2] The value scale can be relativized to the speaker, the listener, or even a third party. Thus, the value scale helps us concretize face considerations; a speaker's effort to save face can be equated to their effort to maximize the subjective value for themselves, the listener, or someone else (Yoon et al., 2016). In this study, we focus on the speaker's face, and thus treat the value scale as associated with the speaker. We assume that the speaker's face is threatened when they make a negative comment because they may be perceived as a negative person. The listener, in turn, takes the value scale into account in inferring what the speaker intends to convey by an utterance.

[3] We will not discuss the entity type in this paper, for this factor did not have any significant effect on ratings.

[4] The interpretation of negated adjectives can be affected by whether the negation marker receives prosodic focus. To control for such an effect, we used a contracted form of negation (*wasn't*) instead of the canonical form (*was not*).

participants based on the speaker's target utterance (*How important was it to* ***speaker*** *to be honest?* and *How important was it to* ***speaker*** *to be positive?*).

There were thus three independent variables: whether the adjective in the target utterance occurred in a non-negated or negated form; whether the adjective had inherently positive or negative polarity; and whether the preceding context of the utterance created a default or reverse value scale. Figure 1.A exemplifies a negated form, positive polarity, reverse value scale condition.

There were a total of 64 unique items (8 adjectives * 2 negation conditions * 2 value scale conditions * 2 entity types). Each participant completed 8 target trials, which consisted of each adjective paired with one of the eight conditions. In addition, participants completed 8 control trials which served the purpose of calibrating their responses and preventing them from guessing that the experiment was about negation. The controls differed from target trials in that the utterance never contained negation and that the speaker directly evaluated the listener's work (e.g. *Paul [...] bakes vegan cupcakes. [...] When Paul asks for feedback, Dora says: "It's tasty."*), using one of 2 antonym pairs (*entertaining/boring, tasty/gross*). The preceding context directly described the true state (desirable, undesirable) instead of the value scale, followed by an utterance containing an adjective (positive, negative). There were 32 control items in total (4 adjectives * 4 items * 2 true states). The order of main and control trials was randomized.

Each participant completed a two-trial practice phase before the main phase of the experiment. In the practice phase, participants were given feedback when they provided incorrect responses to the VALUE QUESTION and the STATE QUESTION. Finally, each participant was asked to complete an optional language background survey.

Results

We conduct three analyses to test the predictions laid out above. We begin with a manipulation check to ensure that our manipulation of the value scale succeeded. Next, we ask two questions based on a single analysis of state ratings: whether the negative strengthening asymmetry replicates; and whether the direction of the asymmetry changes as a result of flipping the value scale. Finally, we examine whether participants infer about the speaker's social vs. informational goals differently when the value scale is flipped.

VALUE and GOAL scale endpoints were coded as 0 (left endpoint) and 1 (right endpoint) for analysis. For the STATE scale, 1 was anchored to the adjective in the utterance (i.e., for an utterance of "Her driving was/wasn't fast," 0 reflected the slow and 1 the fast endpoint of the scale; for "Her driving was/wasn't slow," 0 reflected the fast and 1 the slow endpoint).

Manipulation check Figure 1.B shows mean ratings in response to the VALUE QUESTION (*What does* ***speaker*** *want the* ***entity*** *to be like?*). If the manipulation check succeeded, mean ratings should be high in the default and low in the reverse condition. This result was confirmed by a mixed effects linear regression[5] predicting VALUE ratings from a fixed effect of value scale condition and by-adjective random intercepts: there was a significant main effect of value scale such that participants gave lower ratings in the reverse condition ($\beta = -0.83$, $SE = 0.004$, $t = -178.96$, $p < 0.001$).

State rating analysis Figure 1.C shows mean state ratings by negation, polarity, and value scale condition. To address our two questions of interest, we conducted a mixed-effects linear regression predicting state rating from fixed effects of negation (reference level: 'non-negated'), polarity (reference level: 'positive'), value scale (reference level: 'default'), and their interactions, as well as random by-adjective intercepts and slopes for value scale. Table 1 shows the full model summary.

Table 1. Mixed-effects linear regression results

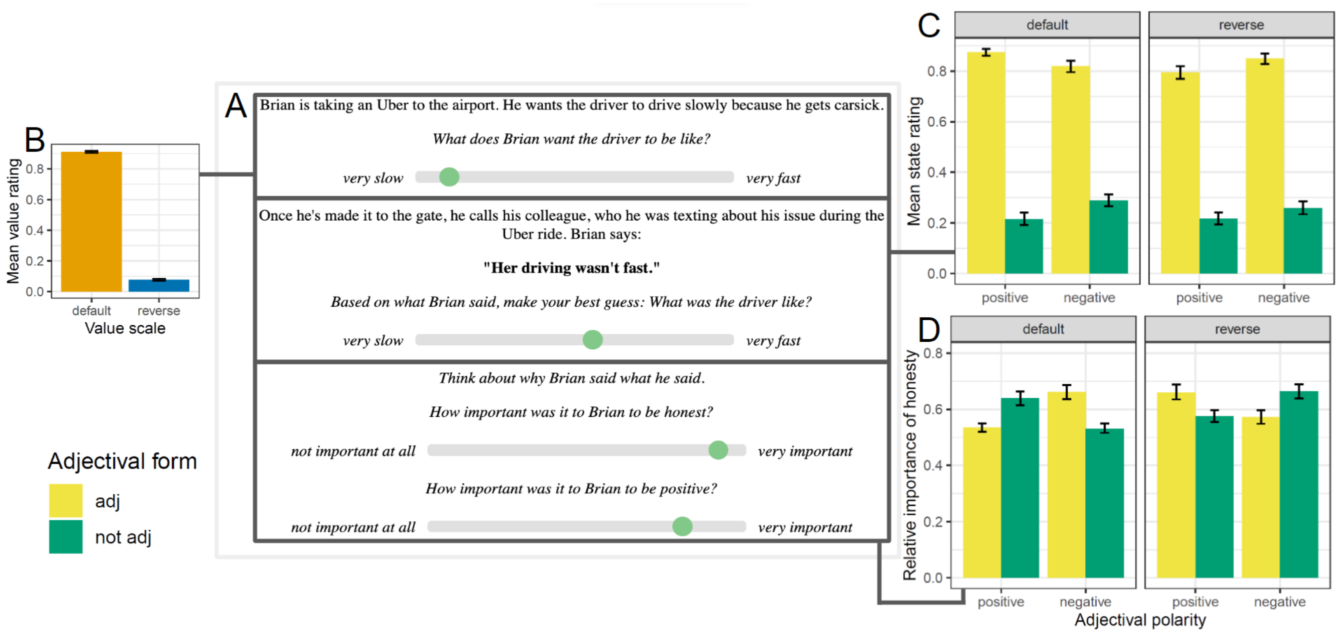
Fixed effect	β	SE	t	$p <$
(Intercept)	0.87	0.01	57.95	0.001
negated	-0.66	0.03	-22.03	0.001
negative	-0.05	0.02	-2.47	0.04
reverse	-0.08	0.03	-2.45	0.04
negated:negative	0.13	0.04	3.00	0.02
negated:reverse	0.08	0.02	3.41	0.001
negative:reverse	0.10	0.04	2.34	0.05
negated:negative:reverse	-0.13	0.03	-4.15	0.001

The asymmetry in negative strengthening was indeed replicated in the default condition, as evidenced in the significant interaction between negation and polarity: the difference between non-negated and negated forms was larger for positive adjectives than for negative ones. Note that this asymmetry was achieved not only through lower ratings of negated positive adjectives vs. their non-negated counterpart (0.22 vs. 0.87) but also through lower ratings of non-negated negative adjectives vs. their positive counterpart (0.82 vs. 0.87). This suggests negative polarity adjectives display *positive weakening*, perhaps due to participants shying away from providing undesirable state ratings. This is a new finding that required studying negated and non-negated forms of adjectives in tandem.

We now turn to our crucial question of interest: whether the negative strengthening asymmetry persisted (as predicted by the Polarity Hypothesis) or flipped (as predicted by the Face Hypothesis) when the value scale was flipped. As evidenced in the significant three-way interaction, the asymmetry changed.

To further investigate this change, we conducted a pairwise comparison of estimated marginal mean differences between

[5] All regression analyses were conducted using the lme4 package in R (Bates et al., 2015), and each regression model included the maximal random effects structure that allowed model convergence without incurring a singularity issue.



Brian is taking an Uber to the airport. He wants the driver to drive slowly because he gets carsick.

What does Brian want the driver to be like?

very slow ● *very fast*

Once he's made it to the gate, he calls his colleague, who he was texting about his issue during the Uber ride. Brian says:

"Her driving wasn't fast."

Based on what Brian said, make your best guess: What was the driver like?

very slow ● *very fast*

Think about why Brian said what he said.

How important was it to Brian to be honest?

not important at all ● *very important*

How important was it to Brian to be positive?

not important at all ● *very important*

Figure 1. A: example trial. B: mean VALUE ratings (manipulation check). C: mean STATE ratings. D: mean HONESTY ratings, normalized.

the negated and non-negated forms of adjectives for both polarity and value scale conditions, using the emmeans package in R (Lenth et al., 2021). The results are shown in Figure 2. In the default condition, the difference between non-negated and negated forms was significantly larger for positive polarity adjectives (*was fast* - *wasn't fast*) than for negative polarity adjectives (*was slow* - *wasn't slow*). This indicates that in the default condition, the degree of negative strengthening was larger for positive than negative adjectives. However, while the relative size of the difference was flipped for the reverse condition – i.e., negative polarity adjectives (*was slow* - *wasn't slow*) had a larger difference between non-negated and negated forms than positive polarity adjectives (*was fast* - *wasn't fast*) – the 95% confidence intervals overlapped substantially. This indicates that the polarity asymmetry was not flipped but rather eliminated in the reverse condition.

Goal rating analysis If participants interpret positive and negative adjectives differently when the value scale is flipped, do they also infer the speaker's goals differently? We indeed see such patterns in Figure 1.D, which reports normalized HONESTY ratings, calculated as HONESTY rating / (HONESTY rating + POSITIVITY rating). In the default condition, the more face-threatening utterances (*slow, not fast*) received higher HONESTY ratings than the more face-saving utterances (*fast, not slow*), indicating an inference toward a lower weight of the social goal upon hearing a face-threatening utterance. This pattern was flipped in the reverse condition, suggesting that participants perceived positive polarity adjectives (*fast*) as conveying a more undesirable state than negative polarity adjectives (*slow*), thus inferring the speaker valued being honest more than being positive when they said "Her driving was fast."

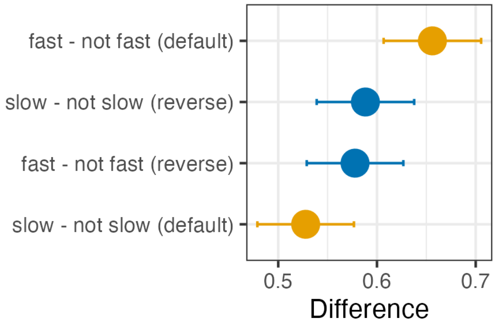


Figure 2. Pairwise differences of the estimated marginal means: non-negated vs. negated adjectives in default (yellow) and reverse (green) value scale condition, for positive polarity (*fast*) and negative polarity (*slow*) adjectives.

Altogether, the results suggest that *both* polarity and face play a role in negative strengthening. This is a novel finding that is not predicted by either the Polarity Hypothesis or the Face Hypothesis on their own. We show below that an RSA model containing components of both polarity and face can account for these results jointly.

RSA modeling

We use the RSA framework (Frank & Goodman, 2012; Goodman & Frank, 2016; Degen, 2023) to model negative strengthening in a unified way, incorporating the effects of both adjectival polarity and face considerations. Specifically, we adopt the polite RSA model of Yoon et al. (2020) and enrich it with a threshold semantics for adjectives (Kennedy, 2007; Lassiter & Goodman, 2017; Tessler & Franke, 2018) and a cost for polarity.

Model specification

The RSA framework models communication as interlocutors recursively and probabilistically reasoning about each other. A pragmatic listener interprets utterances by reasoning about a pragmatic speaker. The pragmatic speaker makes utterance choices by reasoning about a literally interpreting listener.

At the bottom of the recursion, the literal listener L_0 observes an utterance u containing an adjective in either negated or non-negated form. Given a threshold θ associated with the adjective, it maps the utterance to a state s that is above the threshold (if positive polarity) or below the threhold (if negative polarity):

$$L_0(s|u,\theta) \propto [\![u]\!](s,\theta) \cdot P(s) \tag{1}$$

where $[\![u]\!](s,\theta)$ is a meaning function that returns true if the state s passes the threshold θ associated with the utterance u, and $P(s)$ is a uniform prior on states.

$$[\![u]\!](s,\theta) = \lambda\theta\lambda s(s > \theta) \tag{2}$$

Next, the pragmatic speaker S_1, with a state s in mind to communicate, chooses an utterance u that soft-maximizes utterance utility U and minimizes cost $C(u)$[6]:

$$S_1(u|s,\theta,\phi,\beta) \propto \exp\Big\{\alpha \cdot \big[U(u,s,\phi,\beta) - C(u)\big]\Big\} \tag{3}$$

The speaker utility function contains two terms to reflect that the speaker aims to maximize not only the probability that the literal listener will arrive at the intended state s (informational utility U_{info}), but also the subjective value associated with each state (social utility U_{social}). The relative weight of each utility is determined by ϕ:

$$U(u,s,\phi,\beta) = \phi \cdot U_{\text{info}}(u,s) + (1-\phi) \cdot U_{\text{social}}(u,\beta) \tag{4}$$

Here, U_{info} is the amount of information the literal listener would have after hearing an utterance:

$$U_{\text{info}}(u,s) = \ln L_0\big(s|u,\theta\big) \tag{5}$$

and U_{social} is the expected subjective value given an utterance:

$$U_{\text{social}}(u,\beta) = \mathbb{E}_{L_0(s|u,\theta)}\left[V(s,\beta)\right] \tag{6}$$

where V is a value function that maps a state to a subjective value. Given that a state has a measurement on a scale, $V(s,\beta) = s \cdot \beta$, where β represents the size and the direction of the value scale associated with the adjectival scale. The sign of β therefore controls whether the value scale operates in default or reverse mode, and the absolute value of β represents the perceptual importance of the value scale.

We assume the cost of an utterance $C(u)$ in S_1 is modulated by negation and polarity. Thus, $C(u)$ is modeled to vary depending on whether it is a negated or non-negated form, and whether it contains a positive or negative polarity adjective.

Finally, the pragmatic listener L_1, given an utterance u and the value scale β, jointly infers the state s, the threshold θ, and the relative weight of informational utility ϕ based on a model of S_1 and prior beliefs about s, θ, and ϕ:

$$L_1(s,\theta,\phi|u,\beta) \propto S_1(u|s,\theta,\phi,\beta) \cdot P(s) \cdot P(\theta) \cdot P(\phi) \tag{7}$$

Our model differs from its predecessors in several ways. In contrast with Tessler & Franke (2018), we introduce varying polarity costs to the model, as a way of testing the explanatory importance of polarity. In comparison with Yoon et al. (2020), we extend the concept of U_{social} to cover the speaker's face. Depending on whose face is under threat, β can variably represent different value scales. Another difference is that β is not a constant but a variable whose value is determined in context and can be negative. Finally, our meaning function is not based on empirical data but on uncertain thresholds, which are also inferred by the pragmatic listener.

We implemented our model in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). A complete implementation can be found at `https://github.com/sarangjeong/negated-adjectives`.

Model evaluation

To replicate the experimental results, we explored different values of β (size and direction of the value scale) and costs of polarity. We report model predictions when β is set as 5 (corresponding to the default value scale) or -1 (corresponding to the reverse value scale) and the costs of {*was, wasn't, fast, slow*} are set as $\{1,2,0,0.5\}$.[7] The aggregate model predictions are shown in Figures 3.

The model succeeded in replicating the major patterns found in the empirical data. First, the asymmetric negative strengthening in the default condition (cf. Figure 1.C, left facet) was predicted by the model with $\beta = 5$ (expected state of *not fast* = 0.36, expected state of *not slow* = 0.43). Second, when $\beta = -1$, the asymmetry was eliminated (expected state of *not fast* = 0.40, expected state of *not slow* = 0.40), which corresponds to the empirical pattern observed in the reverse condition (cf. Figure 1.C, right facet). Noteworthy is that a model that did not include costs of polarity (the costs of {*was, wasn't, fast, slow*} set as $\{1,2,0,0\}$) predicted a flipped asymmetry in the reverse condition (expected state of *not fast* = 0.42, expected state of *not slow* = 0.39), suggesting that having a cost on polarity in addition to social utility is crucial

[6] The soft-maximization is governed by parameter α – the higher α is, the more utility-maximizing, i.e., rational, the speaker.

[7] The exact values of the cost were set arbitrarily. We expect the model to make similar predictions as long as the cost of negation is higher than the cost of no negation and the cost of a negative polarity adjective is higher than the cost of a positive polarity adjective.

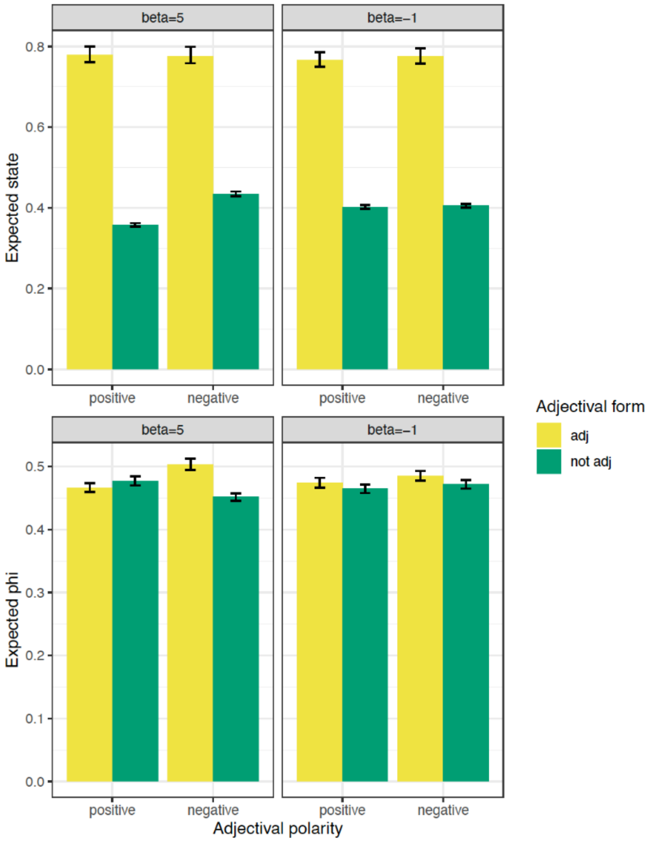


Figure 3. Model predictions for state (top) and ϕ (bottom)

in explaining the empirical patterns observed in the experiment.

The model also made predictions about ϕ, i.e., the speaker's weight on the informational utility relative to the social utility. Model predictions followed the trends in GOAL ratings (cf. Figure 1.D) although it predicted less variability than the empirical patterns and failed to replicate the trend of the negative polarity adjectives in the reverse condition. We speculate this discrepancy between the model prediction and the empirical data comes from the fact that the value scale is parameterized asymmetrically. Namely, β is set as 5 for the default condition, yet it is set as -1 instead of -5 for the reverse condition.[8] A smaller absolute size of β leads to a smaller variance of the social utility, which may have resulted in overlapping values of ϕ.

General Discussion

Why do we interpret "It wasn't good" and "It wasn't bad" asymmetrically? This study explored two possible causes: intrinsic polarity of adjectives and face considerations. To tease them apart, we conducted a behavioral experiment where participants were tasked with inferring the meaning of an utterance given a context that varied in the value scale. In addition, we computationally modeled the behavioral pattern using the RSA framework. The empirical and modeling results suggest that the asymmetry in negative strengthening is the result of both intrinsic adjectival polarity and face considerations.

Empirically, we replicated the polarity asymmetry in negative strengthening. Crucially, we found an effect of the value scale such that a negative adjective showed a higher degree of negative strengthening in a situation where it conveyed a desirable state (reverse condition) in comparison with the typical situation where it conveyed an undesirable state (default condition). However, the asymmetry was removed rather than reversed in the reverse condition, pointing to the coexistence of the polarity and face effects in negative strengthening, in contrast with the results in Mazzarella & Gotzner (2021) that found no effect of face considerations.

Methodologically, our experiment included non-negated forms of adjectives in addition to negated forms while most studies on negative strengthening only used negated adjectives as stimuli (e.g. Mazzarella & Gotzner, 2021; Ruytenbeek et al., 2017).[9] This allowed us to measure negative strengthening directly by comparing the ratings of negated and non-negated forms. This way, we were able to observe the actual size of the asymmetry that would have been unknown if only negated forms had been compared among themselves. In addition, improving upon Mazzarella & Gotzner (2021), we used clearer contexts. We designed the contexts in the reverse condition such that it was natural for the negative polarity adjective to convey the desirable state. The contexts also explicitly mentioned what was the desirable state to make sure the manipulation was taken as intended. We also directly measured the value scale perceived by participants as a sanity check. As a result, we found an effect of the value scale, or face considerations, unlike in Mazzarella & Gotzner (2021).

To our knowledge, this is the first attempt at modeling negative strengthening in the RSA framework. In addition to Yoon et al. (2020)'s social utility, we incorporated the concept of polarity by assigning a different cost to positive and negative polarity adjectives, and the concept of context-dependent value scale by allowing the subjective value function to vary depending on the context. This model was able to replicate key qualitative patterns in the empirical data.

In conclusion, the present study informed the theory of negative strengthening by showing that both polarity and face considerations play a role in the phenomenon and provided a framework for modeling the phenomenon.

[8] Setting β as -5 led the model to predict a flipped asymmetry, which was at odds with the empirical pattern. This suggests that the reduced asymmetry in the reverse condition may stem from the complex nature of face considerations in the condition, namely that a positive polarity adjective conveys an undesirable state in a specific context (e.g., slow driving is desirable when a passenger gets carsick), but it can at the same time convey a desirable state in a general sense (e.g., fast driving is generally a desirable trait of a driver).

[9] Ruytenbeek et al. (2017) did include non-negated forms in Experiment 2, but they only used them as fillers and did not analyze them.

Acknowledgements

We would like to thank Inseong Han, Brandon Papineau, Dingyi Pan, Brandon Waldon, members of the Stanford ALPS Lab, members of Dan Jurafsky's Lab, the audience at the Stanford Linguistics QP Fest, and the audience at the 6th California Meeting on Psycholinguistics for valuable feedback and discussion. We are also grateful to our anonymous CogSci reviewers for their insightful commentary.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Büring, D. (2007a). Cross-polar nomalies. In *Semantics and linguistic theory* (Vol. 17, pp. 37–52).

Büring, D. (2007b). More or less. In *Proceedings from the annual meeting of the chicago linguistic society* (Vol. 43, pp. 3–17).

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological review*, *82*(1), 45.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive psychology*, *3*(3), 472–517.

Colston, H. L. (1999). "not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes*, *28*(3), 237–256.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*. `http://dippl.org`. (Accessed: 2024-1-5)

Gotzner, N., & Mazzarella, D. (2021). Face management and negative strengthening: the role of power relations, social distance, and gender. *Frontiers in psychology*, *12*, 602977.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q-and r-based implicature. *Meaning, form and use in context*.

Horn, L. R. (1989). A natural history of negation.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of verbal learning and verbal behavior*, *10*(3), 244–253.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive psychology*, *8*(4), 441–480.

Kaup, B., & Dudschig, C. (2020, 03). Understanding Negation: Issues in the processing of negation. In *The Oxford Handbook of Negation.* Oxford University Press.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, *30*, 1–45.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. *Presupposition and implicature in compositional semantics*, 163–177.

Lakoff, R. (1973). The logic of politeness: Or, minding your p's and q's. In *Proceedings from the annual meeting of the chicago linguistic society* (Vol. 9, pp. 292–305).

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*, 3801–3836.

Leech, G. N. (1983). *Principles of pragmatics*. Routledge.

Lehrer, A., & Lehrer, K. (1982). Antonymy. *Linguistics and philosophy*, 483–501.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2021). Emmeans: Estimated marginal means, aka least-squares means. r package version 1 (2018). *Preprint at.*

Mazzarella, D., & Gotzner, N. (2021). The polarity asymmetry of negative strengthening: dissociating adjectival polarity from facethreatening potential. *Glossa: a journal of general linguistics*, *6*(1).

Pfister, J. (2010). Is there a need for a maxim of politeness? *Journal of Pragmatics*, *42*(5), 1266–1282.

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *GLOSSA-A JOURNAL OF GENERAL LINGUISTICS*, *2*(1).

Tessler, M. H., & Franke, M. (2018). Not unreasonable: Carving vague dimensions with contraries and contradictions. In *Cogsci.*

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* (pp. 2771–2776).

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 *Cogsci.*

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.