The MIT Press

Polite speech emerges from competing social goals

Erica J. Yoon[1,+]**, Michael Henry Tessler**[1,+]**, Noah D. Goodman**[1]**, and Michael C. Frank**[1]

[1]Department of Psychology, Stanford University

[+]These authors contributed equally to this work.

Keywords: politeness, computational modeling, communicative goals, pragmatics

ABSTRACT

Language is a remarkably efficient tool for transmitting information. Yet human speakers make statements that are inefficient, imprecise, or even contrary to their own beliefs, all in the service of being polite. What rational machinery underlies polite language use? Here, we show that polite speech emerges from the competition of three communicative goals: to convey information, to be kind, and to present oneself in a good light. We formalize this goal tradeoff using a probabilistic model of utterance production, which predicts human utterance choices in socially-sensitive situations with high quantitative accuracy, and we show that our full model is superior to its variants with subsets of the three goals. This utility-theoretic approach to speech acts takes a step towards explaining the richness and subtlety of social language use.

INTRODUCTION

We don't always say what's on our minds. Although "close the window!" could be sufficient, we dawdle, adding "can you please...?" or "would you mind...?" Rather than tell an uncomfortable truth, socially-aware speakers exaggerate ("Your dress looks great!") and prevaricate ("Your poem was so appropriate to the occasion"). Such language use is puzzling for classical views of language as information transfer (Buhler, 1934; Frank & Goodman, 2012; Jakobson, 1960; Shannon, 1948). On the classical view, transfer ought to be efficient and accurate: Speakers are expected to choose succinct utterances to convey their beliefs (Grice, 1975; Searle, 1975), and the information conveyed is ideally truthful to the extent of a speaker's knowledge. Polite speech violates these basic expectations about the nature of communication: It is typically inefficient and underinformative, and sometimes even outright false. Yet even young speakers spontaneously produce requests in polite forms (Axia & Baroni, 1985), and adults use politeness strategies pervasively – even while arguing (Holtgraves, 1997), and even though polite utterances may risk high-stakes misunderstandings (Bonnefon, Feeney, & De Neys, 2011).

If politeness only gets in the way of effective information transfer, why be polite? Most obvious is the fact that we have social relationships to maintain, and most linguistic theories assume speaker behavior is motivated by these concerns, couched as either polite maxims (Leech, 1983), social norms (Ide, 1989), or aspects of a speaker and/or listener's identity, known as *face* (Brown & Levinson, 1987; Goffman, 1967). Face-based theories predict that when a speaker's intended meaning contains a threat to the listener's face or self-image (and potentially the speaker's face), her messages will be less direct, less ef-

ficient, and possibly untruthful. Indeed, when interpreting utterances in face-threatening situations, listeners readily assume that speakers intend to be polite (Bonnefon, Feeney, & Villejoubert, 2009). How this socially-aware calculation unfolds, however, is not well understood. Adopting an example from Bonnefon et al. (2009), when should a speaker decide to say something false ("Your poem was great!" said of an actually-mediocre poem) rather than to tell the truth ("Your poem was bad") or to be indirect ("Some of the metaphors were tricky to understand.")? How do the speaker's goals enter into the calculation?

We propose a utility-theoretic solution to the problem of understanding polite language, in which speakers choose their utterance by attempting to maximize utilities that represent competing communicative goals. Under the classic pragmatic view of language production, speakers want to be informative and convey accurate information as efficiently as possible (Goodman & Frank, 2016; Grice, 1975); this desire for informative and efficient communication we call *informational utility*. In addition, speakers may want to be kind and make the listener feel good (i.e., save the listener's face), for example, by stating positive remarks about the listener. The utility that underlies this goal is a *prosocial utility*.

If a speaker wants to be informative and kind, then she would ideally produce utterances that satisfy both goals. The nuances of reality, however, can make it difficult to satisfy both goals. In particular, when the true state of the world is of low value to the listener (e.g., the listener's poem was terrible), informational and prosocial goals pull in opposite directions. Informational utility could be maximized by stating the blunt truth ("your poem was terrible.") but that would very likely hurt the listener's feelings and threaten the listener's self-image (low prosocial utility); prosocial utility could be maximized through a white lie ("your poem was amazing"), but at the cost of being misleading (low informational utility). In such situations, it seems impossible to be both truthful and kind. A first contribution of our work here is to formalize the details of this tradeoff in order to predict experimental data.

A second contribution of our work is to develop and test a new theoretical proposal. We propose that speakers may navigate their way out of the truth-kindness conflict by signalling to the listener that they care about both of the goals, even while they are genuinely unable to fulfill them. We formalize this notion of *self-presentational utility* and show that it leads speakers to prefer indirect speech: utterances that provide less information relative to alternatives with a similar meaning.

We look at indirect speech in this paper through negated adjectival phrases (e.g., "It *wasn't bad*"). The relationship between negation and politeness is a topic of long-standing interest to linguists and psychologists (Bolinger, 1972; Horn, 1989; Stern, 1931; Stoffel, 1901). Comprehending negation, as a logical operation, can be psychologically more complex than comprehending an unnegated assertion, resulting in difficulty in processing of negations (Clark & Chase, 1972; see Nordmeyer & Frank, 2014 for an underlying pragmatic explanation) as well as failure to recognize or recall the asserted content (Lea & Mulligan, 2002; MacDonald & Just, 1989). Our interest in negation, however, is for its information-theoretic properties: Negating an assertion that has a specific meaning results in a meaning that is less precise and lower in informativity (e.g., negating "Alex has blue eyes" results in the statement that "Alex has eyes that are some color other than blue"). In our paradigm, we use negation as a way of turning a relatively direct statement ("It was terrible") into an indirect statement ("It wasn't terrible") whose interpretation includes some possibilities that

are consistent with or close to the unnegated statement (i.e., the poem was not terrible, but it was still pretty bad).

Multifactorial, verbal theories – like previous proposals regarding politeness – are very difficult to relate directly to behavioral data. Therefore, to test our hypotheses about the factors underlying the production of polite language (what we refer to as its utility structure), we take a model comparison approach. We do this by formalizing the trade-off between different combinations of speakers' utilities in a class of probabilistic models of language use (the Rational Speech Act (RSA) framework; Frank & Goodman, 2012; Goodman & Frank, 2016), with a particular focus on models with and without the self-presentational utility. In this framework, speakers are modeled as agents who choose utterances by reasoning about their potential effects on a listener, while listeners infer the meaning of an utterance by reasoning about speakers and what goals could have led them to produce their utterances. These models build on the idea that human social cognition can be approximated via reasoning about others as rational agents who act to maximize their subjective utility (Baker, Saxe, & Tenenbaum, 2009), a hypothesis which has found support in a wide variety of work with both adults and children (e.g., Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017). Indeed, this class of pragmatic language models has been productively applied to understand a wide variety of complex linguistic behaviors, including vagueness (Lassiter & Goodman, 2017), hyperbole (Kao, Wu, Bergen, & Goodman, 2014), and irony (Kao & Goodman, 2015), among others.

AUTHOR CONTRIBUTIONS

All authors designed research and wrote the paper; E.J.Y. and M.H.T. performed research and analyzed data.
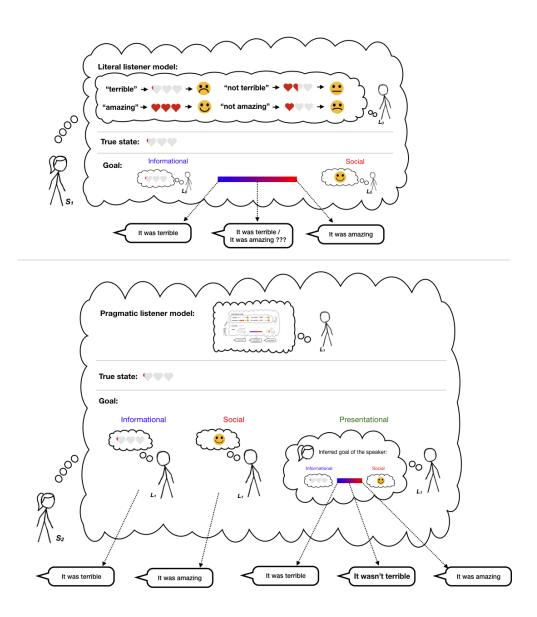
Figure 1. Diagram of the model, showing S_1 (a first-order polite speaker) and S_2 (a higher-order polite speaker capable of self-presentational goals) Top: First-order polite speaker (S_1) produces an utterance by thinking about: (1) the true state of the world (i.e., how good a given performance was); (2) the reasoning of literal listener who updates his beliefs about the true state via the literal meanings of utterances (e.g., not terrible means approximately 1.5 heart out of 3 hearts) and their affective consequences for the listener; and (3) her goal of balancing informational and social utilities. Bottom: Second-order polite speaker (S_2) produces an utterance by thinking about (1) the true state; (2) the pragmatic listener L_1 who updates his beliefs about the true state and the first-order speaker S_1s goal (via reasoning about the S_1 model); and (3) her goal of balancing informational, prosocial, and self-presentational utilities. Different utterances shown correspond to different weightings of the utility components.