Nonliteral understanding of number words

Justine T. Kao[a,1], Jean Y. Wu[b], Leon Bergen[c], and Noah D. Goodman[a]

[a]Department of Psychology and [b]Symbolic Systems Program, Stanford University, Stanford, CA 94305; and [c]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Barbara H. Partee, University of Massachusetts Amherst, Amherst, MA, and approved July 2, 2014 (received for review April 25, 2014)

One of the most puzzling and important facts about communication is that people do not always mean what they say; speakers often use imprecise, exaggerated, or otherwise literally false descriptions to communicate experiences and attitudes. Here, we focus on the nonliteral interpretation of number words, in particular hyperbole (interpreting unlikely numbers as exaggerated and conveying affect) and pragmatic halo (interpreting round numbers imprecisely). We provide a computational model of number interpretation as social inference regarding the communicative goal, meaning, and affective subtext of an utterance. We show that our model predicts humans' interpretation of number words with high accuracy. Our model is the first to our knowledge to incorporate principles of communication and empirically measured background knowledge to quantitatively predict hyperbolic and pragmatic halo effects in number interpretation. This modeling framework provides a unified approach to nonliteral language understanding more generally.

pragmatics | computational modeling

Imagine a friend describing a new restaurant where she recently dined. Your friend says, "It took 30 minutes to get a table." You are likely to interpret this to mean she waited ~30 min. Suppose she says: "It took 32 minutes to get a table." You are more likely to interpret this to mean exactly 32 min. Now, suppose she says: "It took a million years to get a table." You will probably interpret this to mean that the wait was shorter than a million years, but importantly that she thinks it took much too long. One of the most fascinating facts about communication is that people do not always mean what they say—a crucial part of the listener's job is to understand an utterance even when its literal meaning is false. People's ability to interpret nonliteral language poses a critical puzzle for research on language understanding.

A rich body of literature in psychology and linguistics has examined how people use and understand nonliteral language (1–4). However, most of the work has been qualitative, with little focus on analyzing aspects of an utterance that predict the quantitative details of people's figurative interpretations. Here, we present a computational model that formalizes and integrates three general principles of language and communication to explain the basis of nonliteral language understanding. First, speakers and listeners communicate with the assumption that their interlocutors are rational and cooperative agents; second, listeners assume that speakers choose utterances to maximize informativeness with respect to their communicative goals; third, speaker and listener use common ground—their shared knowledge of the world—to communicate effectively. The first principle has been formalized by a recent body of work on rational speech act (RSA) models, which views pragmatic language understanding as probabilistic inference over recursive social models and explains a range of phenomena in human pragmatic reasoning (5–8). We go beyond the previous formal work and address the second principle by extending the RSA framework. We first extend the space of potential interpretations to include subjective dimensions such as affective opinion. We then assume that the listener is uncertain about the speaker's communicative goal and jointly infers both the goal and the intended meaning. Because the interpretation space has multiple dimensions, a speaker's goal may be to maximize the probability of successfully conveying information along one dimension of meaning but not another. This makes it possible for a literally false utterance to be optimal as long as it is informative along the target dimension. These elements of the model have important connections to Gricean pragmatics (9, 10) and relevance theory (11), in particular the argument that listeners infer the meaning of metaphors as well as other forms of loose talk by assuming that speakers maximize relevance (12, 13). Finally, we address the third principle of communication by empirically measuring people's background knowledge to understand the interaction between nonlinguistic and linguistic knowledge in shaping language understanding. By applying this computational approach to a case study on number words, we show that nonliteral interpretations can arise from basic principles of communication without positing dedicated processing mechanisms for nonliteral language.

At the core of RSA models, a listener and a speaker recursively reason about each other to arrive at pragmatically enriched meanings. Given an intended meaning m, speaker S_1 reasons about a literal listener L_0 and chooses utterance u based on the probability that L_0 will successfully infer the intended meaning (7):

$$S_1(u|m) \propto L_0(m|u) \cdot e^{-C(u)}. \quad [1]$$

Here, $C(u)$ is the psychological cost of an utterance, potentially determined by factors such as the utterance's frequency, availability, and complexity. The exponential results from applying a Luce choice rule to model utterance choice, which is used extensively in models of decision making (14). A pragmatic listener L_1 then reasons about S_1 and uses Bayes' rule to infer the meaning m given utterance u, where $P(m)$ is the prior probability of a meaning (although in principle speaker and listener can recurse to arbitrary depth, here we stop at recursive depth 1):

$$L_1(m|u) \propto P(m) S_1(u|m). \quad [2]$$

Because the RSA framework operates under the assumption that speakers optimize informativeness, it predicts that choosing an

Significance

Human communication is rife with nonliteral language, ranging from metaphor to irony to hyperbole. How do people go so far beyond the literal meaning of an utterance to infer the speaker's intended meaning? We present a computational model that understands hyperbolic and other nonliteral uses of number words (e.g., "That watch costs 10,000 dollars"). Our model integrates empirically measured background knowledge, principles of communication, and reasoning about communicative goals to explain the computational basis of nonliteral language understanding. This framework sheds light on the nature of communication, marking a significant advancement in the flexibility and richness of formal models of language understanding.

Author contributions: J.T.K., J.Y.W., L.B., and N.D.G. designed research; J.T.K., J.Y.W., and L.B. performed research; J.T.K. and J.Y.W. analyzed data; and J.T.K. and N.D.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: justinek@stanford.edu.

utterance whose literal meaning directly contradicts the intended meaning is never optimal. However, this contradictory use is precisely the case in nonliteral language. For example, people understand the utterance "It took a million years to get a table" to mean that the wait time was long but not, in fact, a million years, resulting in a contradiction between literal and interpreted meaning. This suggests that the basic RSA model is incomplete and requires additional elements to explain nonliteral communication.

Previous work has examined people's communicative reasons for using figurative language and suggested that certain goals, such as conveying emotion and emphasis, are commonly satisfied by nonliteral language (1). A natural extension is thus to add an affective dimension to the meaning of utterances, which has interesting connections to previous work on expressives (15). However, simply adding this dimension is insufficient; it is still unclear how people infer affect from an utterance whose literal semantics is unconnected to affect (such as number terms). Here, we additionally extend the RSA framework to represent alternative communicative goals, such that a speaker can want to convey information about one dimension but not another. We show that the combination of these two extensions is sufficient to give rise to nonliteral understanding of language.

We explore the case where the interpretation space has two dimensions: the state of the world and the speaker's affect or opinion. In what follows, we describe the subtext dimension as "affect," but it could be other kinds of speaker attitude, *mutatis mutandis*. The speaker is now modeled as follows:

$$S_1(u|s,a,g) \propto \sum_{s',a'} \delta_{g(s,a)=g(s',a')} L_0(s',a'|u) \cdot e^{-c(u)}, \qquad [3]$$

where the intended meaning includes two dimensions s (the state of the world) and a (the speaker's affect). The function g projects the listener's inferred meaning onto relevant dimensions, meaning the speaker's communicative goal is to be informative (only) along this "topic" dimension. A literal listener interprets utterances literally without reasoning about the speaker, whereas a pragmatic listener performs joint inference on both the speaker's goal and her intended meaning:

$$L_1(s,a|u) \propto \sum_{g} P_S(s)P_A(a|s)P_G(g)S_1(u|s,a,g). \qquad [4]$$

The listener uses nonlinguistic background knowledge of the probability of a state (P_S) and the probability of having a particular affect given a state (P_A), which we measure empirically (*Experiment 3a: Price Prior* and *Experiment 3b: Affect Prior*). Based on the listener's linguistic knowledge, the literal semantics of utterance u conveys information about state s and nothing about affect a. However, the common knowledge that affect is usually associated with certain states of the world allows the listener to believe information about a given an assertion about s. If it is known that the speaker's goal is to convey affect, and not the state, then the pragmatic listener will discount information about s but retain information about a—a nonliteral interpretation is obtained. Even when the pragmatic listener is not certain of the speaker's goal, a joint inference of goal, state, and affect can also result in nonliteral interpretation. Common knowledge of a domain and joint reasoning about communicative goals thus allows the speaker to communicate additional dimensions of meaning without explicitly describing these dimensions.

The incorporation of goal inference and multiple dimensions of meaning is a major change to the existing RSA framework that critically allows it to accommodate nonliteral language understanding. As a case study, we focus on the interpretation of number words. We chose number words because they have precise literal meanings that can be easily modeled, and apply to domains (such as prices) that lend themselves to quantitative measurement. We aim to capture two well-known phenomena regarding number interpretation: hyperbole and pragmatic halo. Hyperbole is a figure of speech that uses exaggeration to convey emphasis and emotion (16). Despite being literally false, hyperbolic utterances are readily understood and serve purposes such as establishing social closeness and expressing opinions (1, 16–18). Pragmatic halo refers to people's tendency to interpret round numbers such as 100 imprecisely and sharp numbers such as 103 precisely (19). The halo effect has been formalized in game-theoretic models as a rational choice given different utterance costs and a possibility of pragmatic slack (20, 21). Other research has shown that speakers' tendency to choose simple number expressions decreases when more precise information is relevant to the listener (22), suggesting that higher-level pragmatic considerations such as communicative goals directly impact the production and interpretation of round versus sharp numbers. Our model uses alternative communicative goals coupled with differential utterance costs to model the pragmatic halo effect. We show that our framework for pragmatic inference makes quantitative predictions for both hyperbole and pragmatic halo in the interpretation of number words.

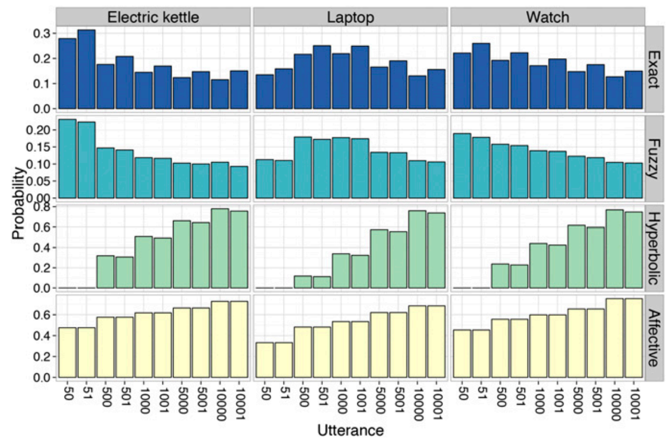

Fig. 1. Model predictions of interpretations given utterances. Each bar in the first three rows shows the probability of a type of interpretation given an utterance. Exact interpretations are more likely given sharp rather than round utterances; fuzzy interpretations are slightly more likely given round utterances; hyperbolic interpretations are more likely given more extreme utterances. The final row shows the probability of an affective interpretation.

Results

We tested our model on number words that refer to the prices of three types of everyday items: electric kettles, watches, and laptops. We selected these items because they have distinct price distributions, P_S, which we measured empirically by asking participants to rate the probability of various prices for the three items (*Experiment 3a: Price Prior*). We also obtained an affect prior, P_A, by asking participants to rate the probability of a speaker thinking that an item is too expensive given a price state (*Experiment 3b: Affect Prior*). Using these priors, which capture purely nonlinguistic knowledge, we aimed to model people's interpretations of utterances such as, "The electric kettle cost u dollars." A formal description of model assumptions is in *Materials and Methods*.

Model Simulations. Using the price priors and affect priors measured for each of the three items, we obtained the meaning distributions predicted by the model for all utterances. Fig. 1 summarizes this distribution into different types of interpretations. The first three are model interpretations regarding the price state: exact (e.g., "1,000" interpreted as 1,000), fuzzy (e.g., "1,000" interpreted as 1,001), and hyperbolic (e.g., "1,000" interpreted as 100). Round utterances (divisible by 10) such as "500" and "1,000" are interpreted less exactly and more fuzzily than their sharp counterparts, which captures pragmatic halo.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

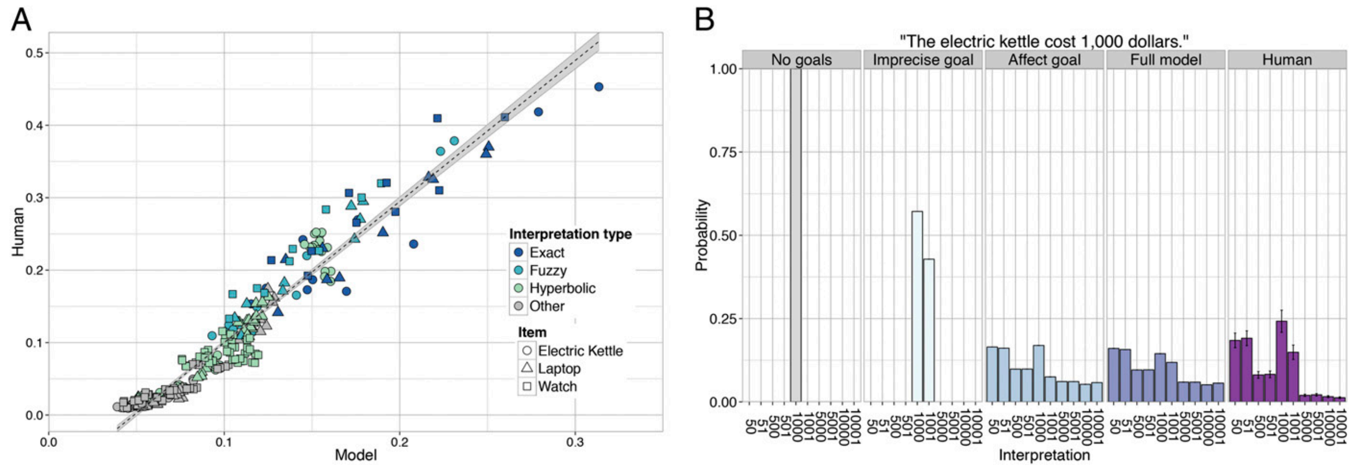

Fig. 2. (*A*) Model predictions vs. average human responses from experiment 1. Each point represents an utterance and price state pair (u, s). The x coordinate of each point is the probability of the model interpreting utterance u as meaning price state s; the y coordinate is the empirical probability. Correlation between model and human interpretations is 0.968 (95% confidence region in gray). (*B*) Comparison of models with different communicative goals and human interpretations for the utterance: "The electric kettle cost 1,000 dollars." A model that considers both affect and precision goals (full model) most closely matches human data.

Utterances whose literal meanings are less likely given the price prior are more likely to be interpreted hyperbolically (e.g., "1,000" is more likely to be interpreted hyperbolically for electric kettles than laptops), which captures a basic feature of hyperbole. Affective interpretation refers to the probability that an utterance conveys the speaker's opinion that the price is expensive. Utterances whose literal meanings are associated with higher affect priors (such as "10,000" and "10,001") are more likely to be interpreted as conveying affect, which predicts the affective subtext of hyperbole.

To build intuition for these predictions, consider a pragmatic listener who reasons about a speaker and analyzes her choice of utterance. The pragmatic listener hears "10,000 dollars" and knows that its literal meaning is extremely unlikely. Given that the speaker reasons about a literal listener who interprets "10,000 dollars" literally and believes that the speaker very likely thinks it is expensive, "10,000 dollars" is an informative utterance if the speaker's goal is to communicate an opinion that the kettle is expensive (without concern for the actual price). Because the pragmatic listener uses this information to perform joint inference on the speaker's communicative goal and the meaning of the utterance, he infers that "10,000 dollars" is likely to mean less than 10,000 dollars but that the speaker thinks it is too expensive.

Behavioral Experiments. We conducted experiment 1 to evaluate the model's predictions for the interpreted price. Participants read scenarios in which a buyer produces an utterance about the price of an item he bought, for example: "The electric kettle cost 1,000 dollars." Participants then rate the likelihood that the item cost s dollars for $s \in S$ (*Experiment 1: Halo and Hyperbole*). Participants were more likely to interpret utterances as hyperbolic when their literal meanings have lower probabilities under the item's prior price distribution [$F_{(1,10)} = 44.06$; $P < 0.0001$]. To examine the halo effect, we computed the difference between the probability of an exact interpretation and the probability of a fuzzy interpretation for each utterance. This difference is significantly smaller for round numbers than for sharp numbers [$F_{(1,28)} = 18.94$; $P < 0.001$], which indicates that round numbers tend to be interpreted less precisely than sharp numbers. To quantitatively evaluate the model's fit, we compared model and human

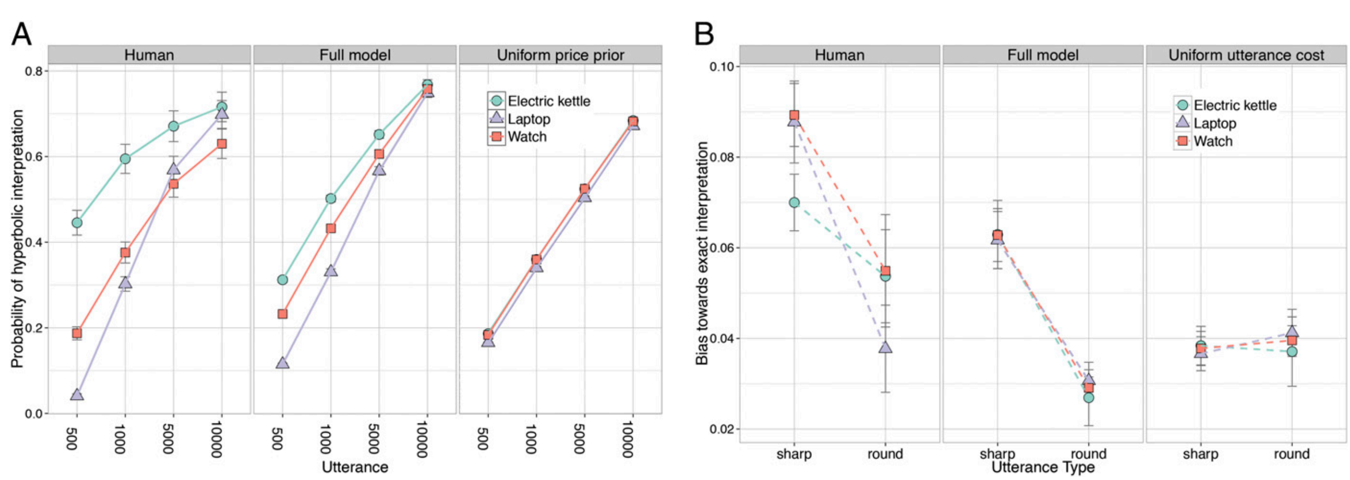

Fig. 3. (*A*) Probability of hyperbolic interpretation given utterances. The leftmost panel shows human data (error bars are SEs). A full model that uses price priors measured in experiment 3a demonstrates similar hyperbole effects and distinguishes among item types; a model that uses uniform price priors does not. (*B*) Halo effect as measure by bias toward exact interpretation for round/sharp utterance types. Humans' bias toward exact interpretation is significantly higher for sharp numbers. A full model that assigns higher cost to sharp numbers captures this result; a model that uses uniform utterance cost does not.

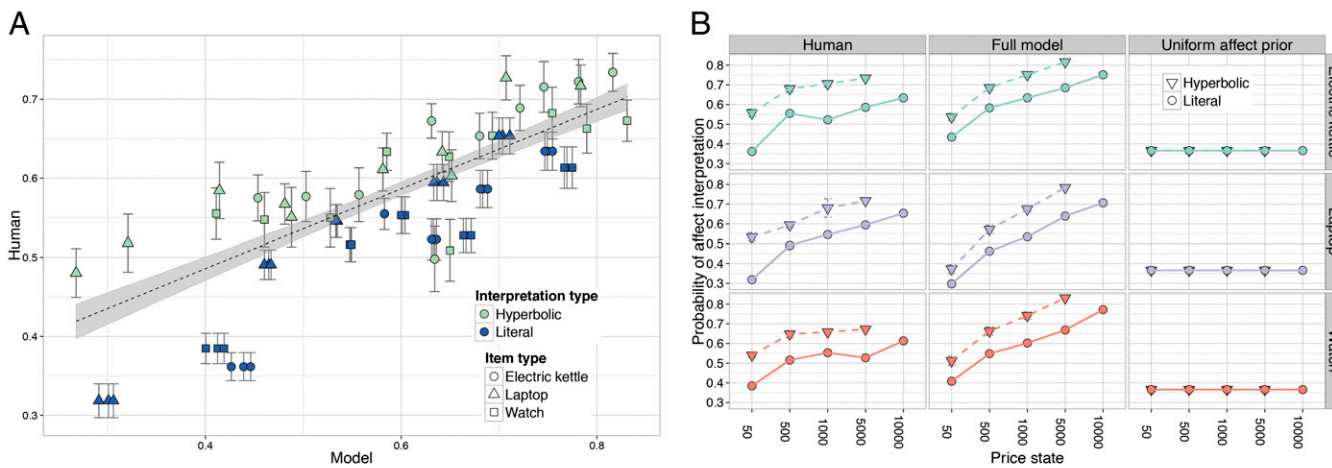

Fig. 4. (*A*) Model predictions of affect vs. human responses from experiment 2. Each point represents an utterance and price state pair (u, s). For pairs where $u = s$, the utterance is literal; for $u > s$, the utterance is hyperbolic. The x coordinate of each point is the model's prediction of the probability that the utterance/price state pair conveys affect; the y coordinate is participants' affect ratings (error bars are SE). Correlation between model and humans is 0.775 (95% confidence region in gray). (*B*) Probability of interpreting a hyperbolic/literal utterance as conveying affect. For the same price state, humans infer higher probability of affect given hyperbolic utterances than literal. A model that uses affect priors measured in experiment 3b captures this result; a model that uses uniform affect priors does not.

interpretation probabilities across all utterances and showed that model predictions are highly correlated with human interpretations of number words ($r = 0.968$; $P < 0.0001$) (Fig. 2*A*; *Materials and Methods* for details).

To show how each component of the proposed model is necessary to capture effects observed in the human data, we explore a series of simpler comparison models. For illustration, Fig. 2*B* compares model interpretations of the utterance "The electric kettle cost 1,000 dollars" given inference over different communicative goals. A model that does not consider alternative goals interprets the utterance entirely literally. Note that even though such a model has information about the affect dimension (i.e., P_A), without goal inference it is unable to produce nonliteral interpretations because it assumes that the speaker only wants to maximize informativeness along the same dimension as the utterance, i.e., the price state. A model that considers a speaker whose goal may be to communicate precisely or imprecisely interprets the utterance as meaning either 1,000 or 1,001. A model that considers a speaker whose goal may be to communicate the price state or her affect prefers price states with higher prior probabilities. Finally, a model that considers the full range of goals demonstrates hyperbole and halo effects that closely match humans' interpretations. To demonstrate that our model is able to usefully incorporate nonlinguistic knowledge to infer the meaning of utterances, Fig. 3*A* shows the hyperbole effect as measured by the probability that an utterance u is interpreted as price state s such that $u > s$. A full model that uses empirically measured price priors captures humans' interpretations, whereas a model that takes a uniform distribution over price states does not. To demonstrate that our model is able to use utterance costs and goal inference to capture pragmatic halo, Fig. 3*B* shows the halo effect as measured by the bias toward exact interpretation for sharp versus round numbers. A full model that assigns a higher utterance cost to sharp numbers captures the significant difference in humans' biases for sharp versus round numbers, whereas a model where utterance costs are uniform does not. These analyses suggest that extending the RSA framework to include goal inference, incorporating empirically measured background knowledge, and including information about utterance costs all contribute to the model's ability to understand nonliteral language.

Does the model capture the rhetorical effect of hyperbole? We conducted experiment 2 to examine humans' interpretation of affect in hyperbolic versus literal utterances. Participants read

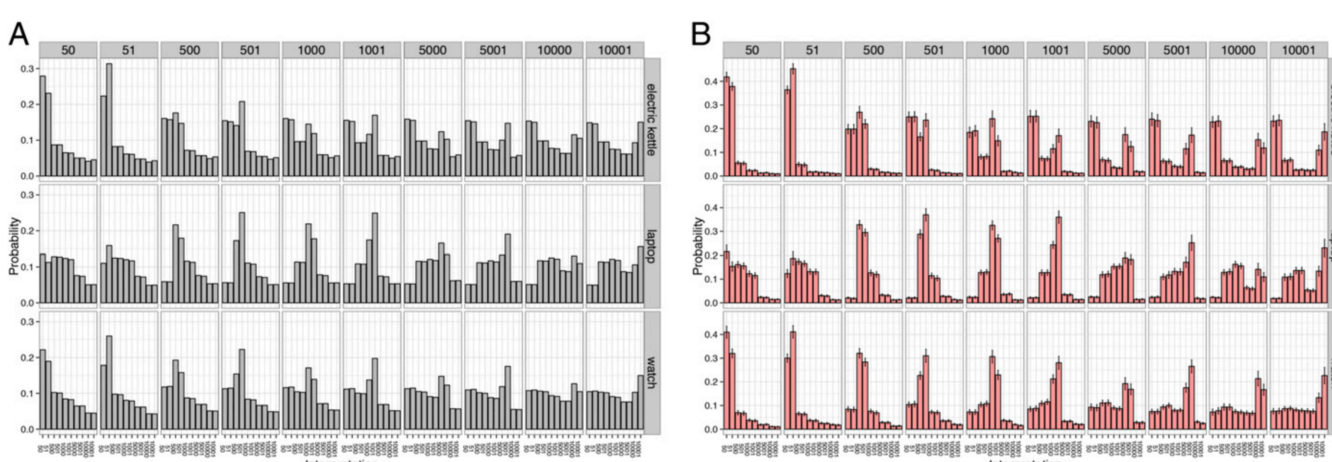

Fig. 5. (*A*) Posterior price state distributions predicted by the model given utterances. Each panel shows the interpretation distribution of an utterance. (*B*) Price state distributions rated by participants given utterances. Each panel shows the interpretation distribution of an utterance. Error bars are SEs.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

scenarios in which a speaker bought an item that cost s dollars and says it cost u dollars, where $u \geq s$. They then rate how likely it is that the buyer thinks the item was too expensive (*Experiment 2: Affective Subtext*). We focused on the affect of an item being too expensive because previous findings suggest that hyperbole is more often used to communicate negative rather than positive attitudes (1, 16). Results showed that utterances u where $u > s$ are rated as significantly more likely to convey affect than utterances where $u = s$ [$F_{(1,25)} = 12.57$; $P < 0.005$]. This suggests that listeners infer affect from hyperbolic utterances above and beyond the affect associated a priori with a given price state. Quantitatively, we compared model and human interpretations of affect for each of the 45 utterance and price state pairs (u, s) where $u \geq s$. Although there is a significant amount of noise in the human judgments (average split-half correlation is 0.833), the model predicts human interpretations of the utterances' affective subtext significantly better than chance ($r = 0.775$; $P < 0.00001$), capturing most of the reliable variation in these data (Fig. 4*A*). To demonstrate how our model explains this effect, Fig. 4*B* shows probabilities of affect given a price state and a literal or hyperbolic utterance. The human data show that higher actual price states are associated with higher probabilities of affect. Within the same price state, hyperbolic utterances are interpreted as conveying more affect than literal utterances. These effects are replicated by the full model, but not by a model that takes in a uniform affect prior. This analysis suggests that the rhetorical effect of hyperbole is driven in part by people's shared knowledge about prices and associated affect.

Discussion

We presented the first (to our knowledge) computational model of nonliteral understanding that quantitatively predicts people's hyperbolic and imprecise interpretations of number words. Our behavioral results show that complex patterns in number interpretation depend on common knowledge between speaker and listener, consideration of communicative efficiency, and, critically, reasoning about the speaker's communicative goal. Our model represents an explicit, computational-level hypothesis about how these factors are integrated to give rise to the particular, graded interpretations at which people arrive. The model's quantitative predictions closely match humans' judgments, including cases of hyperbole, a complex phenomenon previously beyond the scope of computational models.

The current approach has important connections to theories of communication and linguistic meaning. Our speaker aims to be informative, as in Gricean theories of communication, but only with respect to a particular goal or topic—realizing a kind of relevance principle. This relevance is critical for deriving nonliteral interpretations in our model. Although our model is currently limited to two dimensions of meaning and corresponding goals, in future work we hope to capture dimensions central to other figures of speech such as irony and metaphor, thus extending our model to explain nonliteral language more broadly. We believe that our framework significantly advances the flexibility and richness of formal models of language understanding, such that some day probabilistic models will explain everything (hyperbolically speaking).

Materials and Methods

Model. Let u be an utterance. The meaning of u has two dimensions: the actual price state s and the speaker's affect a. We defined the set of price states $S = \{50, 51, 500, 501, 1{,}000, 1{,}001, 5{,}000, 5{,}001, 10{,}000, 10{,}001\}$. We assumed that the set of utterances U is identical to S. We defined the set of affect states $A = \{0, 1\}$ (0 means no affect and 1 means with affect—this binarization is purely for simplicity). Given S and A, the set of possible meanings M is given by $M = S \times A$. We denote each meaning as s, a, where $s \in S$ and $a \in A$.

The speaker S_1 is assumed to be a planner whose goal is to be informative about a relevant topic. We write the goal and its topic as g. S_1 chooses utterances according to a softmax decision rule that describes an approximately rational planner (14):

$$S_1(u|s,a,g) \propto e^{U_1(u|s,a,g)}. \quad [5]$$

We want to capture the notion that the speaker aims to be informative about a topic of discussion while minimizing cost. If the topic is represented by a projection g: $M \to X$ from the full space of meanings to a relevant subspace, then the speaker cares only about the listener's distribution over the subspace,

$$L_0(x|u) = \sum_{s',a'} \delta_{x=g(s',a')} L_0(s',a'|u). \quad [6]$$

Following the RSA model, we formalize informativity of an utterance as the negative surprisal of the intended meaning under the listener's distribution; here, the listener's distribution over the topical subspace X. Hence:

$$U_1(u|s,a,g) = \log L_0(g(s,a)|u) - C(u), \quad [7]$$

where $C(u)$ represents the utterance cost. Substituting into Eq. **5**, this gives the following:

$$S_1(u|s,a,g) \propto \sum_{s',a'} \delta_{g(s,a)=g(s',a')} L_0(s',a'|u) \cdot e^{-C(u)}. \quad [8]$$

In our situations, the speaker may have the goal to communicate along the price dimension, affect dimension, or both. This gives three possible projections r:

$$r_s(s,a) = s$$
$$r_a(s,a) = a$$
$$r_{s,a}(s,a) = s,a.$$

The speaker may also want to communicate the price either exactly or approximately (we assume that no such distinction exists for affect, because we have already binarized it). When the speaker wants to communicate the price approximately, she projects numbers to their closest round neighbors. For example, such a speaker will represent the prices 51 and 1,001 as 50 and 1,000, respectively. This gives two projections (exact and approximate), f, defined as follows:

$$f_e(s) = s$$
$$f_a(s) = \mathrm{Round}(s),$$

where Round(s) denotes the multiple of 10 that is closest to s. The two types of projections, f and r, can be composed to make the goal g of the speaker: $g(s, a) = r(f(s), a)$, which results in $2 \times 3 = 6$ possible goals [although note that $r_a(f_e(s), a)$ and $r_a(f_a(s), a)$ are equivalent].

A literal listener L_0 provides the base case for recursive social reasoning between the speaker and listener. L_0 interprets u literally without taking into account the speaker's communicative goals:

$$L_0(s,a|u) = \begin{cases} P_A(a|s) & \text{if } s = u \\ 0 & \text{otherwise.} \end{cases} \quad [9]$$

The pragmatic listener L_1 performs Bayesian inference to guess the intended meaning given the priors P_S and P_A and his internal model of the speaker. To determine the meaning, the listener will marginalize over the possible goals under consideration:

$$L_1(s,a|u) \propto \sum_g P_S(s) P_A(a|s) P_G(g) S_1(u|s,a,g). \quad [10]$$

The prior probability of s is taken from an empirically derived price prior P_S, and the probability of a given s is taken from an empirically derived conditional affect prior P_A (*Experiment 3a: Price Prior* and *Experiment 3b: Affect Prior*). The probability distribution P_G is defined to be uniform. We used $C(u) = 1$ when u is a round number (divisible by 10) and treated the sharp/round cost ratio as a free parameter that we fit to data (*Experiment 1: Halo and Hyperbole*). We obtained a posterior distribution for all possible meanings s, a given an utterance u. Raw data for model predictions are at http://stanford.edu/~justinek/hyperbole-paper/data/model-predictions.csv. Fig. 5*A* shows the full posterior distributions for all utterances.

Experiment 1: Halo and Hyperbole. A total of 120 participants was recruited on Amazon's Mechanical Turk. We restricted participants to those with IP addresses in the United States (same for all experiments reported). Each participant read 15 scenarios in which a person (e.g., Bob) buys an item (e.g., a watch) and is asked by a friend whether the item is expensive. Bob responds by saying "It cost u dollars," where $u \in \{50, 50 \pm k, 500, 500 \pm k, 1{,}000, 1{,}000 \pm k, 5{,}000, 5{,}000 \pm k, 10{,}000, 10{,}000 \pm k\}$, where k was randomly selected from the set {1, 2, 3} for each trial. We refer to this set of utterances as U. Given an utterance u, participants rated the probability of Bob thinking that the item was expensive. They then rated the probability of the item costing the following amounts of money: 50, 50 $\pm$ k, 500, 500 $\pm$ k, 1,000, 1,000 $\pm$ k, 5,000, 5,000 $\pm$ k, 10,000, 10,000 $\pm$ k, where k was randomly selected from {1, 2, 3} for each trial. We refer to this set of prices as S. Ratings for each price state were on a continuous scale from "impossible" to "extremely likely," represented as real values between 0 and 1. There are a total of 30 possible trial configurations (3 Items $\times$ 10 Utterances). We randomized the order of the trials as well as the names of the buyers (same for all experiments). See stimuli for experiment 1 at http://stanford.edu/~justinek/hyperbole-paper/materials/experiment1.html.

We normalized participants' ratings across price states for each trial to sum up to 1. The average normalized ratings across participants for each item/utterance pair is shown in Fig. 5B, and the data can be found at http://stanford.edu/~justinek/hyperbole-paper/data/experiment1-normalized.csv. To adjust for humans' biases against using the extreme ends of the slider bars, we performed a power-law transformation on the model's distribution: we multiplied the predicted probability for each meaning by a free parameter λ and renormalized the probabilities to sum up to 1 for each utterance. We jointly fit λ and the model's cost ratio C to optimize correlation with the behavioral data. The best fit was with $\lambda = 0.36$ and $C = 1.3$, resulting in a correlation of $r = 0.974$ (95% confidence interval = [0.9675, 0.9793]). The range of cost ratios that produces correlations within this confidence interval is [1.1, 3.7], which is quite broad, suggesting that the overall model fit is not very sensitive to the cost ratio. To further capture the details of the halo effect, we jointly fit λ and C within this range to a measure that is more sensitive to utterance cost: we computed the difference between the probabilities of exact versus fuzzy interpretations for each utterance, which gives us each utterance's bias toward exact interpretation. We then computed the difference in this bias for sharp versus round numbers, which gives us a "halo" score for each sharp/round pair. We fit λ and C to minimize the mean squared error between the model and humans' halo scores. We found that the cost ratio that best captures the magnitude and pattern of the halo effect found in participants' data are 3.4, whereas $\lambda = 0.25$. This produces an overall correlation of 0.9677 with human data from experiment 1. All figures and analyses that we report in the main text are with these parameter values.

For the analysis reported in Fig. 3A, we computed the probability of a participant interpreting an utterance u as hyperbolic by summing up ratings for each interpreted price state s where $u > s$. Because our analysis of hyperbole does not involve utterance costs, we collapsed across round and sharp versions of utterances and price states. For example, "1,001" interpreted as 1,000 does not count as hyperbole. Because 50 and 51 are the lowest available price states, the probabilities for hyperbolic interpretation of utterances "50" and "51" are 0. We computed the average probability of a hyperbolic interpretation across subjects for each utterance. We then showed the hyperbole effect with a linear regression model, using prior probabilities for the utterances' literal meanings as predictor and probabilities for hyperbolic interpretation as response. Results indicated that participants were more likely to interpret utterances as hyperbolic when their literal meanings have lower prior probabilities [$F_{(1,10)} = 44.06$; $P < 0.0001$]. For Fig. 3B, we analyzed the pragmatic halo effect by computing each subject's bias for interpreting an utterance u exactly versus fuzzily. Bias was measured by subtracting the probability of a fuzzy interpretation from the probability of an exact interpretation. We then obtained the average bias for each utterance across subjects. We showed that the average bias for exact interpretation is significantly higher for sharp utterances than for round utterances [$F_{(1,28)} = 18.94$; $P < 0.001$].

Experiment 2: Affective Subtext. A total of 160 participants was recruited on Amazon's Mechanical Turk. Each participant read 30 scenarios in which a person (e.g., Bob) buys an item that costs s dollars and is asked by a friend whether the item is expensive. Bob responds by saying "It cost u dollars," where $u \in U$ and $u \geq s$. Participants then rated how likely Bob thinks the item was expensive on a continuous scale ranging from "impossible" to "absolutely certain," represented as real values between 0 and 1. There is a total of 180 trial configurations (3 Items $\times$ 60 {u, s} pairs, where $u \geq s$). The stimuli for experiment 2 can be found at http://stanford.edu/~justinek/hyperbole-paper/materials/experiment2.html; the raw data at http://stanford.edu/~justinek/hyperbole-paper/data/experiment2-raw.csv. Because our analysis of affective subtext does not involve utterance cost, for the analyses reported in Fig. 4 A and B, we collapsed round and sharp versions of each utterance and price state such that there are a total of 45 utterance/price state pairs under consideration. Utterances u for which $u = s$ are considered literal; utterances u for which $u > s$ are hyperbolic. For the analysis reported in Fig. 4B, we obtained average ratings of affect for each utterance given that it is literal or hyperbolic. A linear regression model showed that hyperbolic utterances are rated as having significantly higher affect than literal utterances across price states [$F_{(1,25)} = 12.57$; $P < 0.005$].

Experiment 3a: Price Prior. To obtain people's prior knowledge of the price distributions for electric kettles, laptops, and watches, 30 participants were recruited from Amazon's Mechanical Turk. Each participant rated the probability of someone buying an electric kettle, laptop, and watch that cost s dollars ($s \in S$), without any linguistic input from the buyer. Ratings for each price state were on a continuous scale from "impossible" to "extremely likely," represented as real values between 0 and 1. The stimuli for experiment 3a can be found at http://stanford.edu/~justinek/hyperbole-paper/materials/experiment3a.html. We normalized participants' ratings across price points for each trial to sum up to 1. The average normalized ratings for each item were taken as the prior probability distribution of item prices. These price distributions were used in the model as P_S to determine the prior probability of each price state. The normalized ratings can be found at http://stanford.edu/~justinek/hyperbole-paper/data/experiment3a-normalized.csv.

Experiment 3b: Affect Prior. To obtain people's prior knowledge of the probability of affect given a price state, 30 participants were recruited from Amazon's Mechanical Turk. Each participant read 15 scenarios where someone had just bought an item that cost s dollars ($s \in S$) without any linguistic input from the buyer. They then rated how likely the buyer thinks the item was expensive on a continuous scale from "impossible" to "absolutely certain," represented as real values between 0 and 1. The stimuli for experiment 3b is at http://stanford.edu/~justinek/hyperbole-paper/materials/experiment3b.html. The average ratings for each price state were taken as the prior probability of an affect given a price state and used in the model as P_A. The data can be found at http://stanford.edu/~justinek/hyperbole-paper/data/experiment3b-raw.csv.

ACKNOWLEDGMENTS. This work was supported in part by National Science Foundation Graduate Research Fellowships (to J.T.K. and L.B.), by a John S. McDonnell Foundation Scholar Award (to N.D.G.), and by Office of Naval Research Grant N000141310788 (to N.D.G.).

1. Roberts RM, Kreuz RJ (1994) Why do people use figurative language? *Psychol Sci* 5(3): 159–163.
2. Dews S, Winner E (1999) Obligatory processing of literal and nonliteral meanings in verbal irony. *J Pragmatics* 31(12):1579–1599.
3. Glucksberg S (2001) *Understanding Figurative Language: From Metaphors to Idioms* (Oxford Univ Press, Oxford).
4. Gibbs R (1999) Figurative language. *MIT Encyclopedia of the Cognitive Sciences*, eds Wilson RA, Keil FC (MIT Press, Cambridge, MA), pp 314–315.
5. Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998.
6. Goodman ND, Stuhlmüller A (2013) Knowledge and implicature: Modeling language understanding as social cognition. *Top Cogn Sci* 5(1):173–184.
7. Bergen L, Goodman GD, Levy R (2012) That's what she (could have) said: How alternative utterances affect language use. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (Cognitive Science Society, Wheat Ridge, CO).
8. Jäger G, Ebert C (2009) Pragmatic rationalizability. *Proceedings of the Sinn und Bedeutung*, eds Riester A, Solstad T (University of Stuttgart, Stuttgart, Germany), Vol 13, pp 1–15.
9. Grice HP (1975) Logic and conversation. *Syntax and Semantics, Volume 3: Speech Acts*, eds Cole P, Morgan JL (Academic, New York), pp 41–58.
10. Clark HH (1996) *Using Language* (Cambridge Univ Press, Cambridge, UK), Vol 4.
11. Sperber D, Wilson D, Ziran H (1986) *Relevance: Communication and Cognition* (Blackwell, Oxford).
12. Wilson D, Carston R (2006) Metaphor, relevance and the "emergent property" issue. *Mind Lang* 21(3):404–433.
13. Sperber D, Wilson D (1985) Loose talk. *Proc Aristotelian Soc* 86:153–171.
14. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
15. Potts C (2007) The expressive dimension. *Theor Linguist* 33(2):165–198.
16. McCarthy M, Carter R (2004) There's millions of them: Hyperbole in everyday conversation. *J Pragmatics* 36(2):149–184.
17. Gibbs RW (2000) Irony in talk among friends. *Metaphor Symbol* 15(1-2):5–27.
18. Gibbs RW, O'Brien J (1991) Psychological aspects of irony understanding. *J Pragmatics* 16(6):523–530.
19. Lasersohn P (1991) Pragmatic halos. *Language* 75(3):522–551.
20. Bastiaanse H (2011) The rationality of round interpretation. *Vagueness in Communication* (Springer, Berlin), pp 37–50.
21. Krifka M (2007) Approximate interpretation of number words: A case for strategic communication. *Cognitive Foundations of Interpretation*, eds Bouma G, Kramer I, Zwarts J (Koninklijke Nederlandse Akademie van Wetenschapen, Amsterdam), pp 111–126.
22. Van der Henst J, Carles L, Sperber D (2002) Truthfulness and relevance in telling the time. *Mind Lang* 17(5):457–466.