Probabilistic pragmatics explains gradience and focality in natural language quantification

Bob van Tiel
Donders Institute for Brain, Cognition and Behaviour, Nijmegen
Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin

Michael Franke
Osnabrück University

Uli Sauerland
Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin

Prefinal version
Published in "Proceedings of the National Academy of Sciences"
https://doi.org/10.1073/pnas.2005453118

Significance statement

Theoretical linguistics postulates abstract structures that successfully explain key aspects of language. However, the precise relation between abstract theoretical ideas and empirical data from language use is not always apparent. Here, we propose to empirically test abstract semantic theories through the lens of probabilistic pragmatic modelling. We consider the historically important case of quantity words (e.g., 'some', 'all'). Data from a large-scale production study seem to suggest that quantity words are understood via prototypes. But based on statistical and empirical model comparison, we show that a probabilistic pragmatic model that embeds a strict truth-conditional notion of meaning explains the data just as well as a model that encodes prototypes into the meaning of quantity words.

Abstract

An influential view in philosophy and linguistics equates the meaning of a sentence to the conditions under which it is true. But it has been argued that this truth-conditional view is too rigid, and that meaning is inherently gradient and revolves around prototypes. Neither of these abstract semantic theories makes direct predictions about quantitative aspects of language use. Hence, we compare these semantic theories empirically by applying probabilistic pragmatic models as a link function connecting linguistic meaning and language use. We consider the use of quantity words (e.g., 'some', 'all') which are fundamental to human language and thought. Data from a large-scale production study suggest that quantity words are understood via prototypes. We formulate and compare computational models based on the two views on linguistic meaning. These models also take into account cognitive factors, such as salience and numerosity representation. Statistical and empirical model comparison show that the truth-conditional model explains the production data just as well as the prototype-based model, when the semantics are complemented by a pragmatic module that encodes probabilistic reasoning about the listener's uptake.

Introduction

An influential tradition in philosophy equates the meaning of a sentence to its *truth conditions* (Frege, 1962; Wittgenstein, 1922). Accordingly, to know what a sentence means is to know in which circumstances it is true or false. This truth-conditional approach to meaning has given rise to the field of *formal semantics*, which uses tools from logic and set theory to formally spell out how the truth-conditional meaning of a sentence is built up from the meanings of its constituents (Heim & Kratzer, 1998).

Other authors have criticized the truth-conditional approach (Lakoff, 1987; Langacker, 1987), drawing upon psychological research that shows that categories are built around *prototypes* that exemplify the most important features of a category (Rosch & Mervis, 1975). The typicality of an exemplar is gradient and depends on similarity to prototypes. The presence of prototype structure is sometimes construed as evidence that category membership—and consequently truth itself—is a matter of degree and depends on similarity to prototypes. This approach is called *Prototype Theory* (PT).

There is an important gap between linguistic meaning and language use, especially from the truth-conditional perspective. For example, many formal semanticists hold that 'or' expresses logical disjunction. However, utterances containing 'or' also tend to imply that the speaker is unsure which disjunct is true (e.g., 'He is in Amsterdam or Berlin'). To explain such observations, Grice outlined a theory of *pragmatics* connecting meaning and use based on the idea that speakers are *rational*, i.e., gear their contributions towards reaching certain conversational goals (Grice, 1975). In the case at hand, a rational speaker who knows which disjunct is true would simply utter that disjunct (e.g., 'He is in Amsterdam') rather than the disjunction (Sauerland, 2004).

Recently, this type of Gricean reasoning has been formalized within probabilistic models that make precise predictions about quantitative aspects of language use (Frank

& Goodman, 2012; Franke & Bergen, 2020; Franke & Jäger, 2016; Potts, Lassiter, Levy, & Frank, 2016; Russell, 2012). Here we leverage these probabilistic pragmatic models to test semantic theories based on empirical data. With this goal in mind, we focus on the specific case of *quantity words*, such as 'some' and 'all'.

Quantity words

Quantification is a crucial aspect of human cognition that supports generalization and underpins the communication of information about frequency and number (Peters & Westerståhl, 2006; van Deemter, 2010). Indeed, the treatment of quantification is one of the centerpieces of the truth-conditional approach to meaning, leading to the development of *Generalized Quantifier Theory* (GQT) (Barwise & Cooper, 1981; Montague, 1973).

GQT

GQT formalizes an intuitive approach to the meaning of quantity words in set-theoretic terms. According to this approach, quantified statements express *relations between sets*. More specifically, many quantity words—and all of the ones that we are concerned with here—express *thresholds* on the *intersection* between two sets (Bonevac, 2012). For example, 'Some S are P' means that the sets S and P denoted by the subject and predicate have at least one element in common, i.e., $|S \cap P| \geq 1$. Other examples of GQ-theoretic definitions are:

(1) a. 'all': $|S \cap P| \geq |S|$
b. 'not all': $|S \cap P| < |S|$
c. 'no': $|S \cap P| \leq 0$
d. 'most': $|S \cap P| > |S - P|$

For convenience, we will refer to $|S \cap P|$ as the *intersection set size*, and to $|S|$ as the *total set size*.

GQT has enjoyed immense theoretical success, explaining various otherwise puzzling observations about, e.g., the cognitive complexity of quantity words (Szymanik & Zajenkowski, 2010), their order of acquisition (Katsos et al., 2016), the aptitude of people's reasoning with quantified sentences (Geurts, 2003), and the distribution of *negative polarity items* like 'any' and 'ever' (Ladusaw, 1979). At the same time, however, it has been observed that there is no straightforward connection between the set-theoretic meanings postulated by GQT and the way people *use* quantity words, as we will show presently.

Using quantity words

We conducted a large-scale study on the production of quantity words in English (Exp. 1a). Each of 600 participants described 10 displays showing 432 circles which were either red or black. To describe these displays, participants freely completed the sentence frame '___ of the circles are red'. Each display varied the intersection set size, i.e., the number of red circles. To elicit quantity words that are potentially vague and thus relevant to our purposes, the displays had a large total set size, thereby likely triggering only approximate representations of the true intersection set size. Additionally, experimental instructions discouraged the use of numerical expressions like 'fifty-two'.

Fig. 1AB visualizes the production probabilities of the 15 most frequently produced quantity words, all of which were produced 50 times or more, as well as 'all' and 'none'. 'All' and 'none' were included because they are crucial for communication and historically salient, even though they were not very frequently produced in the experiment for the obvious reason that displays with uniformly colored dots occurred infrequently. Taken together, these 17 quantity words make up 87% of the production data.

In line with earlier interpretation studies (Newstead, Pollard, & Riezebos, 1987; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986), the results of Exp. 1 suggest that the ranges in which quantity words are produced lack clear boundaries and peak in restricted subintervals of their GQT-meanings. It has been argued that this type of gradience and focality is fundamentally incongruous with GQT (Newstead, 1988; van Tiel, 2014; Zadeh, 2004)—or even with a bivalent truth-conditional approach to meaning more generally (Lakoff, 1987; Langacker, 1987). Instead, it has been argued that quantity words have a prototype-based semantics (Newstead, 1988; van Tiel, 2014); e.g., the prototype for 'most' may be situated at about 75% of the total set size, with the typicality of a situation decreasing with the distance from that prototype.

Outline

The goal of this paper is twofold. First, we show how abstract semantic theories can be compared empirically with data-driven statistical modelling. Second, we show that GQT's truth-conditional account of quantity word meaning offers an equally compelling account of quantity word production as a semantics based on prototypes; but only when it is complemented by a probabilistic theory of language use.

The next section formalizes GQ-theoretic and PT-based semantics of quantity words. We then describe four speaker models that make probabilistic predictions about quantity word choices as a function of the underlying semantic theory.

Semantics of quantity words

A lexical meaning function $\mathfrak{L}\colon M \times T \to [0,1]$ maps each pair of quantity word m and state t to a truth value in the unit interval. Participants in Exp. 1 were presented with

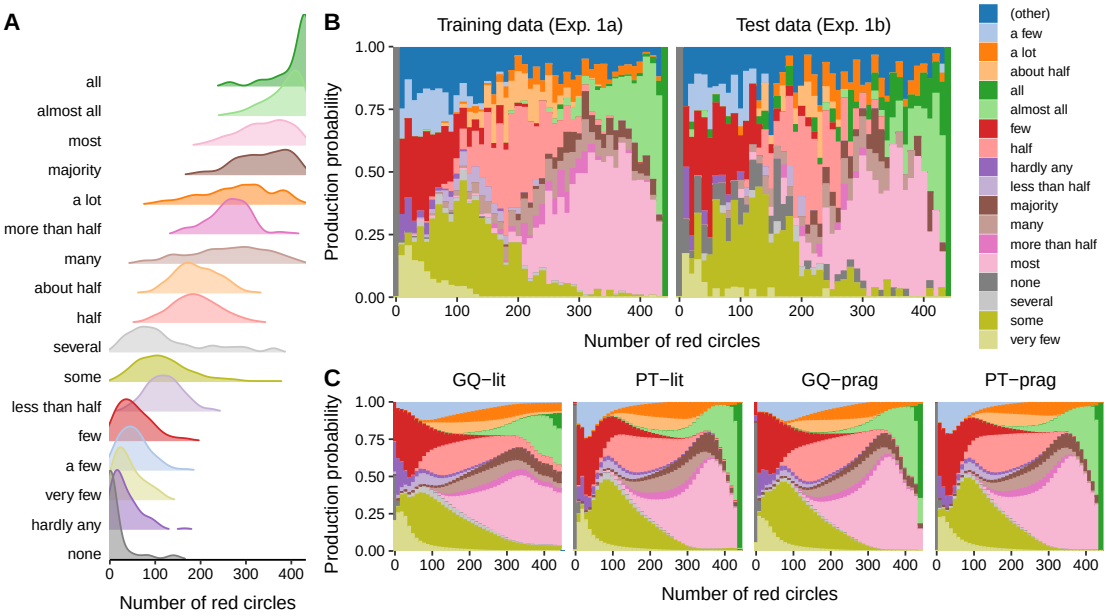

Figure 1: **A**: Density plot showing the distribution of each quantity word over intersection set sizes (Exp. 1a). The height of the distribution indicates the production frequency. **B**: Barplot showing the production probabilities in each bin of intersection set sizes. Each bar shows which quantity words were produced in that bin and their production probability. **C**: Barplot showing the predicted production probabilities for each of the four computational models.

displays consisting of 432 circles which were either red or black. There were thus 433 possible intersection set sizes $T = \{t_0, \ldots, t_{432}\}$, i.e., relevant states of the world.

GQT semantics

According to GQT, many quantity words—and all of the ones that we are interested in—express *thresholds* on the intersection set size $|S \cap P|$ (Bonevac, 2012). Whether a quantity word expresses a lower bound, an upper bound, or both can be judged from the inferential potential in its predicate, i.e., its *right monotonicity* (hence, simply, 'monotonicity').

Quantity words in natural language are either *monotone increasing*, *monotone decreasing*, or *non-monotone* (Barwise & Cooper, 1981). Monotone increasing quantity words license inferences from sets to supersets; monotone decreasing quantity words from sets to subsets; non-monotone quantity words license neither type of inference. Thus, 'all' is monotone increasing, and 'no' is monotone decreasing, as shown by the inference patterns below.

(2) a. All guests ate salmon. → All guests ate fish.
b. No guest ate fish. → No guest ate salmon.

The monotonicity of a quantity word determines the type of threshold that it places on the intersection set size. Monotone increasing quantity words place a *lower* bound; monotone decreasing quantity words place an *upper* bound. The relation between monotonicity and meaning for non-monotone quantity words is less straightforward. However, none of the quantity words in our sample was perceived as non-monotone.

To model the production patterns shown in Fig. 1, we need to make assumptions about the monotonicity properties of the 17 relevant quantity words. We ground these assumptions in empirical data (Exp. 2): 120 participants judged the validity of inference

patterns similar to those shown above. Quantity words were classified as non-monotone if neither argument was accepted more than 50% of the time; otherwise, quantity words were classified based on whether they clustered with 'all' (monotone increasing) or 'none' (monotone decreasing). Following these criteria, 'few', 'hardly any', 'less than half', 'none', and 'very few' were classified as monotone decreasing; all other quantity words as monotone increasing.

Based on these results, we introduce a GQ-theoretic lexicon $\mathfrak{L}_{GQ}(m, t)$, which assigns each pair of quantity word m and intersection set size t the value 1 (true) or 0 (false), based on the threshold θ_m denoted by that quantity word

$$\mathfrak{L}_{GQ}(m, t) = \begin{cases} 1 & \text{if } m \text{ is monotone increasing \& } t \geq \theta_m \\ 1 & \text{if } m \text{ is monotone decreasing \& } t \leq \theta_m \\ 0 & \text{otherwise.} \end{cases}$$

Previous research has shown that people's intuitions about the meanings of quantity words often differ from the textbook definitions, even in the case of 'all' (Newstead & Griggs, 1984), where an imprecise or loose reading could be applied (Krifka, 2002). Therefore, we do not fix the thresholds of quantity words in advance, but treat them as free variables in our model, which are to be inferred from the data. We use Bayesian inference to encode a priori expectations informed by linguistic theory about the likely meanings of quantity words as weakly informative prior distributions over thresholds (Gelman, Carlin, Stern, & Rubin, 2014). (See SI Appendix for more details.)

PT semantics

PT differs from the truth-conditional approach to linguistic meaning in two respects: it assumes that truth is gradient rather than binary, and it holds that linguistic meaning is

organized around prototypes. Consequently, we formalize a PT-based view on quantification using *fuzzy logic*. Where classical logic assumes that sentences must be either true or false, fuzzy logic holds that the truth value of a sentence may take any value in $[0, 1]$ (Zadeh, 1983, 2004).

We thus introduce a prototype lexicon $\mathfrak{L}_{PT}$ that associates quantity words with functions from intersection set sizes to degrees of truth. Each quantity word m has a prototype p_m which is the intersection set size in which the quantity word is maximally true, and a scaling factor d_m which modulates the effect of distance from the prototype on degrees of truth. We assume that degrees of truth decrease exponentially with the quadratic distance from the prototype, so that the truth value of a quantity word m for intersection set size t is:

$$\mathfrak{L}_{PT}(m, t) = exp\left(-\left(\frac{t - p_m}{d_m}\right)^2\right)$$

As before, prototypes p_m and scaling factors d_m will be treated as free variables in data-driven Bayesian inference based on weakly informative priors. Fig. 2A visualizes the meaning of a quantity word in the two types of lexica.

Modelling the production of quantity words

It is commonly assumed that the meanings postulated by semantic theories are part of language users' psychology (Lewis, 1970; Partee, 2001). However, usually, these meanings cannot be observed directly in people's language use. What is needed, then, is a pragmatic theory of language use that can embed various semantic theories to produce probabilistic predictions about the likelihood of production choices under different sets of semantic assumptions. We argue here that recent probabilistic pragmatic models (Frank & Goodman,

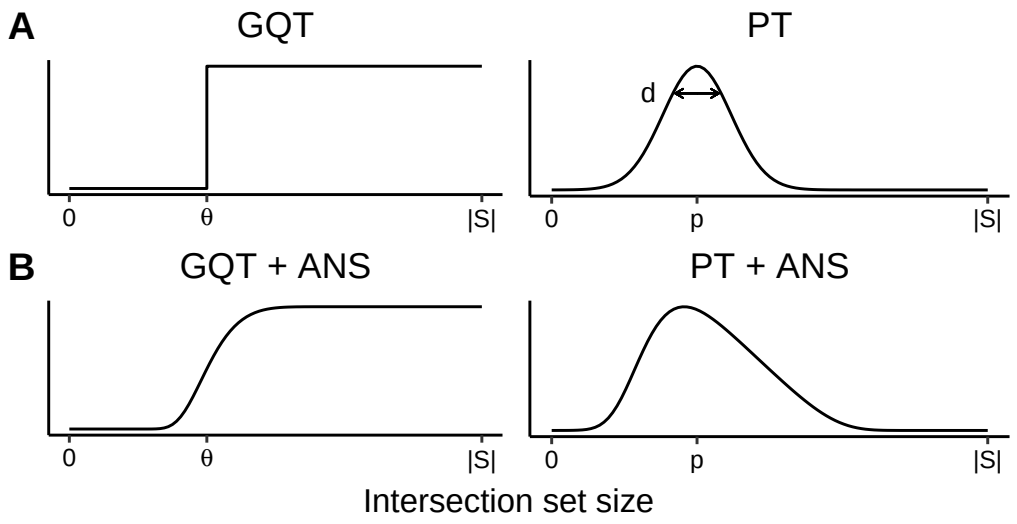

Figure 2: **A**: Visualization of the hypothesized lexical meanings of a (monotone increasing) quantity word in GQT and prototype theory. θ is a threshold value indicating when the quantity word is true or false. p is the prototype meaning, d is the spread around p. **B**: Visualization of the effect of factoring in effects of approximate number representation on the production of the quantity words visualized in Fig. 2A.

2012; Franke & Jäger, 2016) can serve as such an extendable empirical link function. Based on these models, we will be able to infer the most adequate underlying semantics from the observable behavior in experiments based on their causal role in generating the observed behavioral choices (Frank & Goodman, 2014; Schöller & Franke, 2017).

In the following we will therefore introduce two kinds of probabilistic speaker functions. The *literal* speaker preferably produces utterances that are *true*. The *pragmatic* speaker prefers to produce utterances that are both *true* and *informative*. We will compare four models by combining both literal and pragmatic speaker models with GQ-theoretic and PT-based semantics. Each model additionally includes the possibility that some quantity words are more *salient* than others, i.e., come to mind more easily and are therefore more likely to be produced, ceteris paribus. We also factor in *imprecision* in the representation of numerosity.

Literal speakers

A natural starting point for modelling the behavior of participants in a task like Exp. 1 is to assume that, when describing a picture with intersection set size t, participants prefer to give descriptions with a higher truth value over those with a lower truth value, i.e., a preference for true over false descriptions or a preference for descriptions of which t is more prototypical.

To also make room for the idea that not all descriptions come to mind with equal ease, we include the *salience* of quantity words as an additional factor. Participants in the experiment had marked preferences for some quantity words over others. For example, 'most', 'the majority', and 'more than half' are often viewed as truth-conditionally equivalent despite having different conditions of use. Nonetheless, participants used 'most' much more often than 'more than half' or 'the majority'. We assume that this difference in

frequency is due to the fact that 'most' was more salient, i.e., accessible in the task at hand. (See SI Appendix for further discussion.) To capture effects of differential salience, we pair each quantity word with a salience value $P_{\text{Sal}}(m)$, which is treated as a free variable.

Combining truth-based and salience-based production preferences, we consider a simple model of a literal speaker, which serves as a baseline for later comparison:

$$P_{S_{\text{lit}}}(m \mid t, \mathfrak{L}) \propto P_{\text{Sal}}(m)\ \mathfrak{L}(m, t)$$

Pragmatic speakers

One of the central insights of modern pragmatics is that speakers' behavior can be understood, at least in substantial part, as optimal or near-optimal goal-directed action for the purpose of communicating information (Grice, 1975). On this assumption, speakers are predicted to preferably produce utterances not just based on how true they are, but also based on how likely they are to receive the intended interpretation (Frank & Goodman, 2012; Franke & Jäger, 2016). Consequently, a *pragmatic* speaker produces utterances proportionally to how likely they are to convey the intended meaning to a literally interpreting listener (L_{lit}), while also factoring in salience as before:

$$P_{S_{\text{prag}}}(m \mid t, \mathfrak{L}) \propto P_{\text{Sal}}(m)\ P_{L_{\text{lit}}}(t \mid m, \mathfrak{L})^{\alpha}, \text{where}$$
$$P_{L_{\text{lit}}}(t \mid m, \mathfrak{L}) \propto \mathfrak{L}(t, m)$$

This model implies that speakers prefer, all else equal, (i) true descriptions over false ones, (ii) informative descriptions over less informative ones, and (iii) more salient descriptions over less salient ones. The higher the "rationality parameter" α, the more pronounced is the preference for the optimization of communicative success. (See SI Appendix for more details.)

The pragmatic speaker rule is rooted in standard Gricean pragmatic theory but also goes beyond. The preference for informative utterances is reminiscent of how *scalar*

inferences, such as the inference from 'some' to 'some but not all', are usually explained (Horn, 1972). The central idea underlying scalar inferences is that words that stand in an entailment relation may form *lexical scales*. For example, 'all' unilaterally entails 'some', which delivers the lexical scale ⟨some, all⟩. Since efficient speakers are expected to provide as much information as relevant, an utterance with the weaker scalar word 'some' may imply that the speaker believes that the corresponding sentence with the stronger scalar word 'all' is false.

The pragmatic speaker model formulated here generalizes this type of reasoning by assuming that words may compete with each other in spontaneous language use even if they do not stand in a proper entailment relation. For example, consider 'some' and 'few'. These two quantity words do not stand in an entailment relation because both may be true when the other is false (e.g., 'few' is true but 'some' is false when t is zero and 'some' is true but 'few' is false when t equals the total set size). Nonetheless, we propose that pragmatic speakers may, for some intersection set sizes t, consider both 'few' and 'some' as viable utterances, and that they prefer whichever utterance promises to be more reliable in communicating t.

Here, we abstract away from other considerations that may influence speakers' choice of quantity words. For example, 'some' and 'few' also differ in their *argumentative direction*, as shown by the following felicity pattern (Ducrot, 1973):

(3) a. Some people liked the food, which is {good / *bad}.
    b. Few people liked the food, which is {*good / bad}.

The evaluatively neutral nature of our experimental task makes it intuitively unlikely that factors such as argumentativity played an important role in our production study.

Imprecise number representation

Participants in Exp. 1 were not told how many circles were red but rather had to estimate the intersection set size based on the displays. The cognitive module used to estimate the numerosity of large sets is the *Approximate Number System* (ANS) (Dehaene, 1997). The accuracy of ANS estimates decreases as the numerosity to be estimated increases. This effect is a corollary of *Weber's law*, which states that the probability of confusing two stimuli increases with their relative rather than absolute difference (Fechner, 1860).

In order to model the accuracy of participants' estimates of the intersection set size t, we consider the following simple definition of the confusion probability $P_{Cf}(t' \mid t)$ of representing the true intersection set size t as t'. Since the visual displays in Exp. 1 were upper-bounded, $P_{Cf}(t' \mid t)$ is defined as the product of the probability $P_{ANS}(t' \mid t)$ of maintaining an approximate representation of the number t as t' and the inverse probability $P_{ANS}(432 - t \mid 432 - t')$. These probabilities, in turn, are specified following common assumptions in the literature on imprecise representation of number:

$$P_{Cf}(t' \mid t) \propto P_{ANS}(t' \mid t)\ P_{ANS}(432 - t' \mid 432 - t)$$
$$P_{ANS}(t' \mid t) = \int_{t'-0.5}^{t'+0.5} \text{Gaussian}(x, \mu = t, \sigma = w\ t)\text{d}x$$

The parameter w stands for *Weber's fraction*, which represents participants' sensitivity to relative differences between intersection set sizes. To prevent the possibility that different production models leverage the Weber fraction's power as a free parameter unduly, we collected independent data to estimate w based on the same kind of stimuli we used in Exp. 1. To this end, Exp. 3 asked 20 participants to estimate the proportion of red circles in the displays used in Exp. 1. Each participant provided estimates for 24 displays using a continuous slider. Based on the results of this experiment, we determined the

maximum likelihood estimate of the Weber fraction, resulting in $\hat{w} = 0.576$, and use this value for all production models reported here. (See SI Appendix for more information.)

Adding imprecise number representation to the previously defined speaker models helps explain gradience in language use. There are certainly other factors that generate gradience (Krifka, 2002), but adding imprecise number representation is motivated given the stimuli used in Exp. 1 and sufficient to capture probabilistic gradience in production behavior. If $P_S(m \mid t, \mathfrak{L})$ is a speaker production rule, either literal or pragmatic, the production probabilities under an approximate representation of the true world state are:

$$P_S^{CF}(m \mid t, \mathfrak{L}) \propto \sum_{t' \in T} P_{Cf}(t' \mid t)\ P_S(m \mid t, \mathfrak{L})$$

Fig. 2B illustrates the effect of factoring in approximate number representation on the production of the quantity words. The ANS component asymmetrically increases the amount of gradience present in the semantics, reflecting the fact that estimates of the intersection set size are less accurate in the middle range than around the two extremes. Note that, implemented in this way, speakers factor in effects of imprecise number representation in an essentially egocentric way, i.e., they do not take into consideration potential biases in the listener's estimates of numerosity.

Model comparison

All four models were implemented in Stan (Stan Development Team, 2018) to obtain samples from the posterior distribution over free parameter values conditioned on data from Exp. 1a. (See SI Appendix for details). We note that, given the conceptual differences between semantic structures, the priors on the semantic parameters (thresholds for GQ models, prototypes and scaling factors for PT models) are necessarily slightly different as well. Therefore, the models compared here are, strictly speaking, the Bayesian

Model	Log-lik (Exp. 1b)	Rating (Exp. 4)	(95% CI)
GQ-lit	-1717	-2.25	(-3.30, -1.30)
PT-lit	-1660	-0.99	(-1.97, -0.00)
GQ-prag	-1625	-0.77	(-1.82, 0.14)
PT-prag	-1675	-1.41	(-2.37, -0.41)

Table 1: Model comparison results. Posterior expected log-likelihood of the test data from Exp. 1b (Log-lik) and mean of difference between ratings for quantity words observed in the data and quantity words predicted by trained models (Rating) with bootstrapped CIs (95% CI).

models including the informative prior specification we used. (See SI Appendix for more discussion).

Fig. 1C shows the *posterior predictive distribution* of the models. The figure suggests that GQ-lit offers a fair approximation of the data from Exp. 1a, but is substantially worse than the corresponding PT-lit model. However, the difference in model fit seems to disappear once the models are enriched with a pragmatic module. For proper statistical model comparison, we look at how well each model, after being trained on data D^{train} from Exp. 1a, is able to predict independent test data D^{test} from Exp. 1b, which replicated Exp. 1a with 200 participants. Table 1 reports the *posterior expected log-likelihood* of D^{test} for model M with its free parameters Θ_M (higher is better):

$$\int P(\Theta_M \mid D^{\text{train}}, M)\ \log P(D^{\text{test}} \mid \Theta_M, M)\ \mathrm{d}\Theta_M\,.$$

We find that GQ-prag offered the best fit to the new production data, substantially better than PT-lit, PT-prag, and ultimately GQ-lit, in that order.

For additional empirical model comparison, we asked 200 participants to evaluate the predictions of each model (Exp. 4). Participants saw the displays used in Exp. 1, and adjusted a slider to judge the adequacy of descriptions of the form 'Q of the dots are red' for a given intersection set size t. The quantity words Q were either sampled from the production data of Exp. 1a for t or from the four models' posterior predictive distribution for t. The average rating (on a 0–100 scale) for the quantity words from the data was 74.6. Table 1 shows how much lower the quantity words produced by the four models were rated. The GQ-prag model was the only one whose predictions were not rated as significantly worse than the data (β = -0.82, SE = 0.45, t = -1.81, p = .07; all other models: $p < .02$). (See SI Appendix for details.)

Taken together, these results show that the GQT-based model provides a compelling account of patterns of gradience and focality in the production of quantity words when combined with a probabilistic theory of rational communication. Indeed, the pragmatic GQT-based model is at least as compelling as a more flexible competitor that directly encodes gradience and focality into the semantics of quantity words.

General discussion

This paper addresses a long-standing debate about the nature of linguistic meaning. Many philosophers and linguists argue that knowing what a sentence means involves knowing when it is true and when it is false. However, this truth-conditional approach, and specifically its assumption of bivalence, has been criticized as overly rigid and ignoring the fact that meaning is inherently gradient and organized around prototypes (Newstead, 1988; Wallsten et al., 1986; Zadeh, 2004).

What makes it challenging to test such semantic theories is that it is difficult, if not impossible, to divorce language from the way it is used (Clark, 1996). Hence, rather

than attempting to evaluate semantic theories directly, we evaluated them based on how well they explain language use when combined with a set of independently justified linking assumptions that connect meaning and use. We showed that the growing field of *probabilistic pragmatics* can provide such link functions, thereby leveraging tools and techniques from probabilistic modelling in the cognitive sciences for the study of abstract theories of meaning from linguistics (Franke & Bergen, 2020; Potts et al., 2016).

We applied this novel approach to the domain of quantity words (e.g., 'some', 'all'). By doing so, we showed that the truth-conditional approach to quantification, i.e., Generalized Quantifier Theory (GQT), is able to account for gradience and focality in the production of quantity words just as well as a more flexible prototype-based semantics.

This conclusion ties in with more theoretical concerns that have been voiced about the foundational ideas of Prototype Theory (Kamp & Partee, 1995; Osherson & Smith, 1981). While these concerns have caused support for PT to wane, its central tenets (i.e., that truth is a matter of degree, and that linguistic meaning revolves around prototypes) are still defended (Hampton, 2007; Novák, 2008). All of these approaches attempt to encode facts about the way people use language into linguistic meaning. We have shown that a modular view, whereby language production consists of a semantic module that calculates the truth-conditional meaning of an utterance, and a pragmatic module that reasons about the probability that the utterance receives the intended interpretation, can explain gradience and focalization in production just as well as a PT-based approach. Within this modular approach, gradience emerges from limitations in perception and rationality, and prototype structure is an epiphenomenon of the tendency to contrast competing messages to achieve optimal communication. In the latter respect, the current argument runs parallel to the debate about the meaning of color words (Gibson et al., 2016).

Our endeavor has an important precursor in research on categorization. Category theorists have asked whether categories are defined on the basis of all-or-none rules or

by means of gradient similarity to prototypes. More recently, it has been shown that the apparent tension between these two approaches disappears within a Bayesian framework that allows for varying degrees of certainty about which rules apply in a given situation (Tenenbaum, 1999). In a similar way, we show that quantity words have crisp meanings but that gradience may emerge as a consequence of competition between viable messages—as well as cognitive and perceptual biases.

Our study tested a number of vague quantity words, such as 'few' and 'many'. Vagueness is a classical topic of debate in philosophy. Several theories are currently salient. Both *contextualism* and *epistemicism* argue that statements with vague words are always true or false simpliciter (Kamp, 1981). Others argue that statements with vague words can have truth values in between 0 and 1 (Goguen, 1969). One argument seemingly in favor of the latter approach is that there is gradience in the use of vague words. However, we have shown that it is possible to explain such gradience as a natural by-product of cognitive limitations in non-linguistic domains—in this case, imprecise number representation.

Recent years have seen a rising interest in probabilistic modelling of language use (Frank & Goodman, 2012; Franke & Jäger, 2016; Potts et al., 2016; Russell, 2012). This paper builds on that work and shows how such models can be flexibly enriched to include non-linguistic cognitive components, such as the Approximate Number System. We conclude that probabilistic pragmatic modelling can be a strong bridging instrument in future work to build integrated cognitive models of language use that speak directly to established linguistic theory.

Materials and methods

The SI Appendix provides more information about the experiments, analyses, and modelling, and can be found alongside the data, analysis files, and modelling code at https://osf.io/hsytk/. Exp. 1a received ethical approval from the Stanford Non-Medical IRB (10833). All other experiments received ethical approval from the IRB of the German Linguistics Society (DGfS 01.08.2014). Participants received information about their rights and about the purpose of the study. Afterwards, they gave informed consent.

Exp. 1: Production

600 (Exp. 1a) and 200 (Exp. 1b) participants were drafted on Amazon's Mechanical Turk (AMT). Only workers with an IP address from the United States were eligible for participation. In addition, participants were asked in which language they usually counted, and were excluded if they did not answer English. (These constraints hold for all experiments.) Items showed displays containing 432 black or red circles which were randomly scattered across a 25×36 grid. Participants saw a random selection of 10 displays. At the bottom of each display, participants were asked 'How many of the circles are red?' Participants had to complete a sentence of the form '___ are red' by typing freely into the blank. For the analyses, we grouped together synonyms. For the purpose of data visualization (Fig. 1BC), we binned the data into bins consisting of 10 intersection set sizes, except for the first and last bins, which only contained data for intersection set sizes 0 and 432. Moreover, the second to last bin contained 11 intersection set sizes, also including data for intersection set size 431.

Exp. 2: Monotonicity

120 participants were drafted on AMT. Based on a pretest (Exp. 2a), 17 predicate pairs ⟨P1, P2⟩ were selected such that participants agreed that P1 entailed P2, e.g., ⟨play poker, play cards⟩. Using these predicate pairs, two types of arguments were randomly generated for each quantity word:

(4) a. Q of the people P1. → Q of the people P2.
b. Q of the people P2. → Q of the people P1.

Participants indicated whether these arguments were valid. Two annotated examples illustrated the notion of validity.

Exp. 3: ANS

20 participants were drafted on AMT. Displays were as in Exp. 1. Participants saw a random display from each of 24 bins, each including 18 intersection set sizes, and estimated the proportion of red circles by moving a slider.

Exp. 4: Evaluation

200 participants were drafted on AMT. Each saw 20 random displays from Exp. 1. Displays were described by sentences of the form 'Q of the circles are red', where Q was obtained by first sampling an arbitrary number $0, \ldots, 432$ and then sampling from the posterior predictive distribution of each model, as well as from the data observed in Exp. 1. We only tested items where at least two of the sampled expressions differed. Participants rated the adequacy of descriptions by adjusting a slider bar (with endpoints labelled 'bad' and 'good'). Ratings were recalibrated as differences from the ratings for the quantity words

sampled from the data. A mixed effects linear regression model predicting differential rating on the basis of the source of the quantity word (using the data as reference category) with random intercepts for participants, quantity words, and intersection set sizes was the maximal converging model.

Acknowledgements

We thank R. Bååth for contributing to the design of Exp. 1 and giving permission to make use of the data. This research was funded by the German Research Council (DFG FR 3482/2-1, KR 951/14-1, SA 925/4-1, SA 925/11-1, SA 925/17-1) in part within SPP 1727 (Xprag.de), and by the Dutch Science Organisation (Gravitation grant 'Language in Interaction', 024.001.006). We thank B. Geurts, G. Jäger, G. Scontras, and K. Syrett for valuable feedback.

References

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Bonevac, D. (2012). A history of quantification. In D. Gabbay & J. Woods (Eds.), *Handbook of the history of logic (vol. 11)* (pp. 63–126). Elsevier.

Clark, H. H. (1996). *Using language*. Cambridge University press.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Ducrot, O. (1973). *Le preuve et le dire*. Maison Mame.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf & Hartel.

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

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Franke, M., & Bergen, L. (2020). Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, *96*, 77–96.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*, 3–44.

Frege, G. (1962). *Grundgesetze der arithmetik*. Georg Olms.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd

ed.). Chapman.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*, 223-251.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., … Conway, B. R. (2016). Color naming across languages reflect color use. *Proceedings of the National Academy of Sciences*, *114*, 10785–10790.

Goguen, J. A. (1969). The logic of inexact concepts. *Synthese*, *19*, 325-373.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics (vol. 3). Speech acts* (pp. 41–58). Academic Press.

Hampton, J. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*, 355–384.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell.

Horn, L. R. (1972). *On the semantic properties of logical operators in english* (Unpublished doctoral dissertation). University of California, Los Angeles.

Kamp, H. (1981). The paradox of the heap. In U. Mönnich (Ed.), *Aspects of philosophical logic* (pp. 225–277). Springer.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K., Hrzica, G.,. . . Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*, 9244–9249.

Krifka, M. (2002). Be brief and vague! And how bidirectional optimality theory allows for verbosity and precision. In D. Restle & D. Zaefferer (Eds.), *Sounds and systems.*

Studies in structure and change: A festschrift for Theo Vennemann (pp. 439–458). Mouton de Gruyter.

Ladusaw, B. (1979). *Polarity sensitivity as inherent scope relation* (Unpublished doctoral dissertation). University of Texas, Austin.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago University Press.

Langacker, R. (1987). *Foundations of cognitive grammar (vol. I). Theoretical prerequisites.* Stanford University Press.

Lewis, D. (1970). General semantics. *Synthese*, *22*, 18–67.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language* (p. 221-242). Reidel.

Newstead, S. E. (1988). Quantifiers as fuzzy concepts. In *Fuzzy sets in psychology* (pp. 51–72). Elsevier.

Newstead, S. E., & Griggs, R. A. (1984). Fuzzy quantifiers as an explanation of set inclusion performance. *Psychological Research*, *46*, 377–388.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*, 178–182.

Novák, V. (2008). A formal theory of intermediate quantifiers. *Fuzzy Sets and Systems*, *159*, 1229–1246.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35-58.

Partee, B. (2001). Montague grammar. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 9995–9999). Pergamon.

Peters, S., & Westerståhl, D. (2006). *Quantifiers in language and logic.* Oxford University Press.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*, 755–802.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Russell, B. (2012). *Probabilistic reasoning and the computation of scalar implicatures* (Unpublished doctoral dissertation). Brown University, Providence, RI.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391.

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*. *Linguistics Vanguard*, *1*, 20160072.

Stan Development Team. (2018). *The Stan core library.* (Version 2.18.0) doi: http://mc-stan.org

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science*, *34*, 521–532.

Tenenbaum, J. (1999). *A Bayesian framework for concept learning* (Unpublished doctoral dissertation). MIT, Boston.

van Deemter, K. (2010). *Not exactly: In praise of vagueness.* Oxford University Press.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*, 147–177.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*, 348–365.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus.* Harcourt.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, *9*, 149–184.

Zadeh, L. A. (2004). Precisiated natural language (PNL). *AI Magazine*, *25*, 74–92.