On the Optimality of Vagueness: "Around","Between" and the Gricean Maxims*

Paul Égré Benjamin Spector Adèle Mortier Steven Verheyen

Abstract

Why is our language vague? We argue that in contexts in which a cooperative speaker is not perfectly informed about the world, the use of vague expressions can offer an optimal tradeoff between truthfulness (Gricean Quality) and informativeness (Gricean Quantity). Focusing on expressions of approximation such as "around", which are semantically vague, we show that they allow the speaker to convey indirect probabilistic information, in a way that gives the listener a more accurate representation of the information available to the speaker than any more precise expression would (intervals of the form "between"). We give a probabilistic treatment of the interpretation of "around", and offer a model for the interpretation and use of "around"-statements within the Rational Speech Act (RSA) framework. Our model differs in substantive ways from the Lexical Uncertainty model standardly used within the RSA framework for vague predicates.

*Acknowledgments. This research was supported by the following projects: ANR PROBASEM (ANR-19-CE28-0004-01), ANR FrontCog (ANR-17-EURE-0017), as well as by ANR-DFG project COLAFORM, and the van Gogh Project (42589PM). We thank audiences in Stuttgart (DGFS congress 2018), Groningen, Amsterdam, Aix-en-Provence, Brussels (LNAT 4), Berlin (Workshop "The Meaning of Numerals"), Stockholm, Turin (FEW 2019), London (COLAFORM Meeting), Nantes (OASIS 2), Munich. Special thanks go to David Atkinson, Mira Ariel, Alexandru Baltag, Leon Bergen, Denis Bonnay, Emmanuel Chemla, Christoph Hesse, Barteld Kooi, Christian List, Peter Pagin, Jeanne Peijnenburg, Philippe Schlenker, Stephanie Solt, Hendrik De Smet, Jan-Willem Romeijn.

1 Introduction

Why is ordinary language vague? Traditional accounts of vagueness generally insist that vagueness is a deficiency, compared to what an ideal language would look like. Russell (1923) defined vagueness as a one-many relation between an expression and its meaning. In an ideal language, the relation would be one-one. And indeed, artificial languages typically eliminate ambiguity and vagueness by the same token.

Various explanations have been proposed to rationalize the vagueness of ordinary language, however. Russell himself made a central observation in connection to language use when he noted: "*it would be a great mistake to suppose that vague language must be false. On the contrary, a vague belief has a much better chance of being true than a precise one, because there are more possible facts that would verify it*" (p. 91). Phrased in Gricean terms, Russell's observation may be put as follows: a cooperative speaker who is not perfectly informed about the world would too often flout the Gricean maxim of Quality if compelled to use only precise expressions. The maxim of Quality says that one should not say what one believes to be false, and that one ought not say that for which one lacks adequate evidence (Grice 1967). Conversely, vagueness may help cooperative speakers to remain both truthful and justified in their assertions (see Égré and Icard 2018).

The thought, although phrased differently, underlies several accounts of vagueness, drawing attention to the relation between vagueness and error-reduction in the face of uncertainty (see Channell 1985; van Deemter 2009). For Channell, "vagueness may be used as a safeguard against being later shown to be wrong" (p. 17). Similarly, van Deemter points out that "the doctor is uncertain how the future will turn out, which is why he, sensibly, wraps his predictions in vagueness" (p. 622).

The idea that the function of vague language is to help speakers comply with the maxim of Quality is however not straightforward, given the availability of semantically precise but logically weak sentences that make it easy to satisfy the maxim of Quality even in cases where one has little information. For instance, a sentence such as *There were between 2 and 97 guests at the party yesterday* has intuitively precise truth-conditions but expresses a logically weak statement, which can be used truthfully by a speaker who knows fairly little. In contrast, an utterance of a vague sentence such as *There were about 45 guests at the party yesterday* intuitively suggests that the speaker has more information than one who utters the former precise sentence.

In this paper, we will argue that certain vague expressions allow for an optimal tradeoff between the maxims of Quality and Quantity. They sometimes allow the speaker to achieve a communicative effect that no semantically precise sentence could. More specifically, focusing on expressions of approximation such as "around", we will argue that such expressions allow the speaker to indirectly convey probabilistic information, so as to comply with the maxim of Quality while achieving high informativity.

The probabilistic dimension of the interpretation of vague expressions has antecedents in the literature. Frazee and Beaver (2010) argue that gradable terms like "tall" or "many" are vague in so far as they constrain "some measure relative to a value which cannot be known in principle or in practice". On their approach, and in agreement with standard theories of the context-sensitivity of gradable expressions (see Kennedy 2007), "tall" semantically means "taller than t", and "many" means "more than m", but speaker and hearer are typically uncertain about those threshold values t and m. Frazee and Beaver's picture of communication, which we endorse in this paper, is that the "information conveyed by a vague sentence is a statistical distribution" over values and thresholds, which interlocutors try to convey to each other. They argue that the use of vague language is rational in situations of uncertainty, and our own proposal is, from this respect, close in spirit to theirs. Lassiter and Goodman (2017) offer a probabilistic model of the pragmatics of vague predicates within the Rational Speech Act model of pragmatics (Goodman and Stuhlmüller 2013). In their model, when interpreting a sentence such as *she is tall*, the listener updates her belief state (viewed as a probability distribution about possible states of the world) in a way that factors in uncertainty about thresholds. The communicative effect of the sentence can then be viewed as the way it affects this posterior belief state. We should note however, that Lassiter and Goodman's (2017) model is one where the speaker is viewed as maximally informed about the variable of interest (say, someone's height), and so does not by itself address the link between vagueness and speaker's uncertainty (we discuss Lassiter and Goodman's 2017 approach in more details in Section 8).

In this paper, we focus on the meaning, use and interpretation of expressions of numerical approximation such as "around" and "about". We will offer a model where such vague sentences end up communicating a probability distribution. More specifically, the speaker, though not fully informed, is assumed to have more information than the listener about some variable of interest, and the sentence used is informative to the extent that the hearer's posterior distribution over world states after processing an utterance is closer to that of the speaker than prior to the utterance. We argue that the reason why speakers may choose a vague statement as opposed to a precise (but logically weak) one is that this allows them to achieve some communicative effects that would not have been achievable by using a precise statement. In particular, coming back to Russell's intuition, vague language might allow speakers to be both informative and truthful even when their epistemic state does not categorically rule out any particular state of affairs, by allowing them to indirectly convey that they take some state of affairs to be more likely than others.

Our first goal is to identify a range of contexts in which the use of "around" is optimal compared to any lexical alternative that would be more precise (section 2). Our second goal is to advance the understanding of the probabilistic semantics of "around" by giving specific attention to the comparison between numerical expressions of the form "around n" and the use of precise intervals of the form "between i and j" (sections 3 and 4). The basic treatment we give of "around" in section 3 is fundamentally listener-oriented, however. In sections 5-7 we explain how to integrate this model within the framework of the Rational Speech Act model of Goodman and Stuhlmüller (2013) (RSA for short), in particular to explain how a pragmatic speaker is going to take into account the listener's probabilistic representation of the competition involving "around" and "between". In section 8, we propose a more detailed comparison with the model of gradable adjectives proposed by Lassiter and Goodman (2017), and, more generally, the Lexical Uncertainty Model of Bergen et al. (2016). Specifically, in our model, the choice between an "around"-message and a "between"-message can depend on the *shape* of the speaker's probability distribution over the variable of interest: the "around-n"-message is prefered by a speaker whose distribution favors values close to n. In contrast, in the Lexical Uncertainty model, keeping all other parameters constant, the speaker's choice of a message only depends on the *support* of her distribution, not its shape – a result proved in Appendix A. In section 9 we outline further potential developments. Appendix B describes a variant of our Bayesian model, originally our first model, making fundamentally identical qualitative predictions, but distinct quantitative predictions. Appendix C presents two alternative models briefly discussed in section 8.

2 When vague is better than precise

When is it rational for a cooperative speaker to use vague as opposed to more precise language?

One class of situations concerns cases in which the speaker is fully knowledgeable and has precise information at her disposal. She may prefer to use vague language, however, if she expects a precise figure to convey irrelevant information to the listener. For instance, to use an example from Veltman (2001), suppose the question under discussion is how fast you can run the steeplechase. The speaker may prefer to say "I can run the steeplechase very fast" than to utter "I can run the steeplechase in 11 minutes 12 seconds", if she expects the listener to not have the slightest idea of racing times in relation to the steeplechase. By using the vague predicate "very fast", the listener can get more efficient information about the speaker's relative position compared to other runners than if communicated absolute temporal information.[1]

Another class of situations, which will be our main focus in this paper, concerns cases in which the speaker herself fails to have precise information at her disposal. Such cases loom large in Williamson (1994)'s epistemic account of vagueness, and they are described by van Deemter (2009) as cases of necessary vagueness. One of Williamson's central examples concerns a subject watching a crowd, and unable to make an exact count of the people in the crowd. Suppose the speaker is attending a party involving a crowd of 77 people, but does not know that number. Imagine that, after the party, the speaker is asked: "how many people were at the party?". By assumption there is no number n for which the speaker can respond: "there were exactly n persons at the party", on pain of violating Grice's maxim of Quality. For either the speaker would fail to say something she thinks is true, or she may by luck give the correct answer, but she would fail to have adequate evidence for it.

How then should the speaker respond? Our account is grounded in the assumption that situations of that kind fundamentally involve a probabilistic representation of the situation. In practice, the speaker has a probability distribution on possible values of the number of invitees at the party. Let us assume, for the sake of the argument, that the speaker knows with certainty that there were no more than 100 invitees, and knows with certainty that there were at least 40 people present. By assumption, the support of the speaker's probability distribution is the interval $[40, 100]$. We may suppose moreover that her distribution has a peak at 70, and that 80% of the probability mass lies between the values 60 and 80 on either side of that peak (see Figure 1).

Assuming the speaker has access to her distribution, the speaker has several strategies in response to the question. One is to communicate the support of her distribution. By uttering "there were between 40 and 100 persons" the speaker is guaranteed to satisfy the maxim of Quality. Intuitively, however, the speaker reports poorly on her information state. Although the proposition "there were between 40 and 100 persons at the party" is the most informative proposition she believes with certainty in response to the question, the information communicated gives no hint concerning the values the speaker deems more likely than others.

Another option would be for the speaker to communicate a narrower interval. For example, she may choose to respond "between 60 and 80", if she thinks that a .8 chance of getting the right result is high enough. This time, the speaker communicates a more informative proposition, still likely to include the true value. Nevertheless, in case 85 people were at the party, the speaker

[1] Interlocutors typically assess vague expressions relative to implicit standards. See for example Verheyen et al. (2018) for empirical evidence that "tall" and "heavy" applied to humans are ascribed in part relative to one's own height and weight.

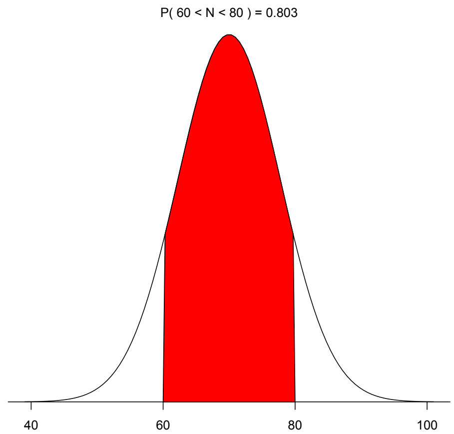

Figure 1: Hypothetical probability distribution on number of people present at the party

runs the risk of ruling out the actual value and of misleading the listener. A case of this kind is a potential violation of the maxim of Quality, or of Williamson's norm of assertion (assert only what you know).

A third option would be for the speaker to communicate her confidence level: "I believe with 80% confidence that there were between 60 and 80 persons at the party". However, that proposition is complex to articulate. Moreover, it forces the speaker to say something explicit about her epistemic state.[2]

Our main claim in this paper is that by uttering "there were around 70 people at the party", the speaker is avoiding the pitfalls of each of the previous options. Even if 85 or 55 people turned up, the speaker is not ruling out either possibility by saying "around 70". In that sense, the utterance is safe and more likely to respect Quality. Secondly, "around 70" intuitively convey something about the shape of the speaker's distribution: 70 should be taken to be more probable to the speaker than other values, and so this message conveys probabilistic information that the "between-sentence" does not. Thirdly, by saying "around 70" the speaker does not have to communicate an explicit confidence interval.

In what follows, we propose to substantiate those intuitions. In order to do so, we proceed to specify a model of the interpretation of "around" in the next section.

[2]In practice, accessing one's confidence level may be a delicate matter too. One might argue that even reports of confidence levels are challenges to the Knowledge norm of assertion. However, our argument above does not rely on that premise, we may assume that the speaker has reliable access to her confidence levels.

3 Modeling "around"

3.1 Around vs. Between

"Around" is sometimes interpreted as specifying a fixed interval whose extension depends on the granularity of a contextually given measurement scale. According to Krifka (2007) and Solt (2014), "around n" denotes the interval $[n - \frac{u}{2}, n + \frac{u}{2}]$, where m is the unit setting the relevant granularity. For instance, "around 10" could denote the interval $[9, 11]$, or $[9.5, 10.5]$, or $[5, 15]$, and so forth, depending on the context.

We agree that the meaning of "around n" should be cashed out in terms of intervals centered on n, but we believe it is inadequate to specify this meaning in terms of a unique fixed interval. Even as the granularity is known to all speakers, say is equal to 10, "around 20" need not be interpreted rigidly as meaning "between 15 and 25". To wit, a speaker who fails to know how many people were at a party and to whom intervals of ten units set the order of magnitude would not necessarily speak falsely by reporting "around 20" if the actual number of attendees were 26 or 27.

We thus see two main differences in the comparison of "around" and "between". First of all, "around" is *semantically vague*, whereas "between" is not. This means that "around n" does not specify a sharp interval; it is compatible with an open-ended range of values, unlike "between". Consider the following two reports:

(1) There were around 70 people at the party.

(2) There were between 60 and 80 people at the party.

Intuitively, if we learn that there were 87 people at the party, (2) appears false stricto sensu, unlike (1). Of course, the utterer of (2) may be using "between" with some slack, and a charitable listener may deem (2) close enough to the truth to be acceptable. However, the point is that for "around" the vagueness in question is directly part of the meaning of the expression. Further confirmation of the vagueness of "around" is given by modification of the target numerals with "exactly". This modification is permitted with "between" but produces gibber with "around":

(3) a. ??There were around exactly 70 people.
    b. There were between exactly 60 and exactly 80 people.

Relatedly, it can be observed that "around n" is sorites-susceptible in a way that "between i and j" isn't. For us, this means that if "k is around n" is considered true, then "k' is around n" is also likely to be judged true when k' is close enough to k but a little more removed from n (Cobreros et al. 2012; Égré et al. 2019). For example, if 19 is around 30, then it seems that 18 is also around 30. But if 20 is between 20 and 30, 19 is not between 20 and 30. For "between" we thus expect the membership function to be a step function, but for "around" we expect a smooth function.[3]

The second main difference we see when comparing "around n" and "between i and j" is more subtle, but will occupy central stage in the rest of this paper. We call it *peakedness*. It concerns the representation of how probable the values are in the interval $[i, j]$ specified by "between", compared to those in neighbourhoods of n in the case of "around". Assume you have no idea how many people, within a certain range, will attend the next evening lecture at the university. You ask the organizer how many people she expects. Compare the following answers:

[3]On sorites-susceptibility and the need for smooth membership functions, see also Borel (1907); Smith (2008); Égré and Barberousse (2014).

(4) a. Between 20 and 40.
b. Around 30.

Our intuition is that (4)-b conveys that the closer a value is to 30, the more likely it is deemed by the speaker in this case. In particular, (4)-b conveys that 30 is more likely to the speaker than other values. By contrast, (4)-a does not appear to convey that any value in the range $[20, 40]$ is more likely than any other: no peakedness results in this case.

The intuition in question is subtle here. We think it will be particularly clear in contexts where, before processing the sentence, the listener does not have strong expectations as to how many people will turn up. This will translate into the assumption that the listener has a uniform prior on the number of attendees, within a certain range. When the prior is not uniform, so when some values are initially more expected than others, peakedness remains in play, as we discuss in section 4, but it may be less manifest.

3.2 A semantics for "around"

In order to derive the previous facts, we propose a Bayesian model of the interpretation of "around". The model is actually a variant of a distinct model that we first came up with and that will be presented later. The two models agree in their main predictions, but an advantage of the Bayesian model is that its conceptual motivation is very clear. The model has two components: a *semantic* component which specifies the meaning of *around*, and an *inferential component* which describes how listeners update their beliefs, when they accept a sentence, on the basis of its meaning. Importantly, the model we propose in this section is listener-oriented. That is, we first account for the effect that using "around" is producing on the listener. We consider the speaker's perspective in section 5.

3.2.1 The model

We assume, following the spirit of a number of former proposals, the truth-conditional meaning of "x is around n" is that $x \in A^n_y$, where A^n_y is an interval of the form $[n-y, n+y]$ and y is an open semantic parameter:

(5) $[\![\text{around}]\!]^y = \lambda n.\lambda x.x \in [n-y, n+y]$

From a purely semantic point of view, then, an utterance of the form "x is around n", cannot express a proposition unless a value for y is provided. However, even if no specific value for y is provided, the listener nevertheless learns something from such an utterance, namely the fact that, whatever the value of y is, x in in the interval $[n-y, n+y]$. If the listener has some expectations about the values that y could take, then she can gain information regarding x. This is, in essence, the idea that our model of the listener will capture. That is, the listener's task is to infer what values x is likely to have, given some uncertainty on what values y is also likely to have. The key point will be that, under uncertainty about the length of the intended interval, a value closer to n is more likely to fall in that intended interval. In this respect, our proposal is close in spirit to Lassiter and Goodman's (2017) approach to gradable adjectives: the taller Mary is, the more likely it is that her height is above the treshold for *tall*, so when learning that Mary is tall, the listener shifts her probability distribution over Mary's height to higher values.[4]

We represent the listener's information state by a joint probability distribution P_L over the possible values taken by x and y. $P_L(x = k)$ represents the prior probability that x takes on a specific value k, and $P_L(y = i)$ represents the prior probability that the interval picked by "around" has radius i. We assume that the prior probabilities of these two types of events are independent, so that in general:

$$P_L(x = k, y = i) = P_L(x = k) \times P_L(y = i) \tag{i}$$

Let n be the number used in 'x is around n'. The information gained by the listener is that x is in the interval $[n - y, n + y]$, where both x and y are random variables. We assume that the goal of the listener is to infer the correct value of x. That is, the listener's problem is to figure out the conditional probability distribution defined by

$$P_L(x = k \mid x \text{ is around } n) = P_L(x = k \mid x \in [n - y, n + y]) \tag{ii}$$

We have:

$$P_L(x = k \mid x \text{ is around } n) = \sum_i P_L(x = k, y = i \mid x \in [n - y, n + y]) \tag{iii}$$

Let us abbreviate $d(x, n) \leq y$ for $x \in [n - y, n + y]$, which is equivalent to $|n - x| \leq y$. Bayes Theorem gives us:

$$P_L(x = k, y = i \mid d(x, n) \leq y) = P_L(d(x, n) \leq y \mid x = k, y = i) \times \frac{P_L(x = k, y = i)}{P_L(d(x, n) \leq y)} \tag{iv}$$

Note that $P_L(d(x, n) \leq y \mid x = k, y = i)$ equals 1 if $d(k, n) \leq i$, and is 0 otherwise. Let $\mathcal{I}$ be defined as a function which takes an arithmetic statement and returns its truth-value. Then we have:

$$P_L(x = k, y = i \mid d(x, n) \leq y) = \frac{\mathcal{I}(d(k, n) \leq i) \times P_L(x = k, y = i)}{P_L(d(x, n) \leq y)} \tag{v}$$

The denominator does not depend on either k or i. Let us call it D to ease calculations.[5] We therefore have:

[4] One significant difference is that in Lassiter and Goodman's (2017) proposal, which is couched in the Rational Speech Act model, the joint reasoning about the variable of interest – say someone's height – and the parameter of interpretation (e.g., a threshold for *tall*) does not take place at the level of the 'literal listener', but is carried out by the first-level pragmatic listener. This difference will prove to have important consequences when we develop a model for the speaker. See section 8 and Appendix A for a detailed discussion.

[5] $D = \sum_k P_L(x = k) \times \sum_{i \geq |n-k|} P_L(y = i)$ — this is the sum of all terms that can be obtained from the numerator by instantiating x with all its possible values.

$$P_L(x = k \mid x \text{ is around } n) = \sum_i \frac{\mathcal{I}(d(k,n) \leq i) \times P_L(x = k, y = i)}{D} \tag{vi}$$
$$= \frac{1}{D} \sum_{i \geq |n-k|} P_L(x = k) \times P_L(y = i)$$
$$= \frac{1}{D} \times P_L(x = k) \times \sum_{i \geq |n-k|} P_L(y = i)$$

In the remaining of this paper, we will often use the *proportionality notation*, whereby the above equation is expressed as follows:[6]

$$P_L(x = k \mid x \text{ is around } n) \propto P_L(x = k) \times \sum_{i \geq |n-k|} P_L(y = i) \tag{vii}$$

3.2.2 Illustration

To illustrate the predictions of this model, let us assume that the listener's prior distribution is uniform both on the values that x might have, and on the interval values y that "around" might denote. Upon hearing "around n", we may assume that $[0, 2n]$ is the largest interval compatible with the meaning of "around n" (if 0 is the scale minimum), hence that $[0, n]$ is the range of values that y can take. For such uniform priors, it follows analytically from Eq. (vi) that:

$$P_L(x = k \mid x \text{ is around } n) = \frac{n - |n-k| + 1}{(n+1)^2} \tag{viii}$$

Consider for instance the effect of a Listener hearing "x is around 20". The posterior distribution of the Listener on the possible values of x is depicted by the histogram (black lines) in Figure 2. The red line represents the uniform prior on the values x might take. As the figure makes clear:

(i) For "x is around 20" the posterior distribution is symmetric and centered on 20, and the further away a value is from 20, the less probable it has become.

(ii) Compare hearing "x is between 10 and 30'. In the latter case, the Listener does a simpler Bayesian update assigning zero probability to the values outside the interval $[10, 30]$. The solid blue line represents the posterior for "between 10 and 30", which is a uniform posterior.

[6] The symbol $\propto$ reads 'is proportional to'. More specifically, a statement of the form $f(a|\ldots) \propto g(a,\ldots)$ is short for: $f(a|\ldots) = \frac{g(a,\ldots)}{\sum\limits_{a' \in A} g(a',\ldots)}$, where A is the domain over which the variable a ranges. So the formula in Equation (vi) boils down to:

$$P_L(x = k \mid x \text{ is around } n) = \frac{P_L(x = k) \times \sum\limits_{i \geq |n-k|} P_L(y = i)}{\sum\limits_{j \text{ is in the support of } x} P_L(x = j) \times \sum\limits_{i \geq |n-j|} P_L(y = i)}$$

.

The proportionality factor ensures that the sum of all $P_L(x = \ldots \mid x \text{ is around } n)$, across all possible values for x, is 1, so that $P_L(x = \ldots \mid x \text{ is around } n)$ is a probability distribution.

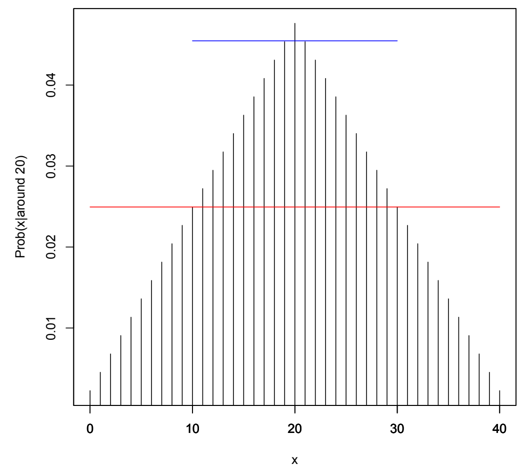

Figure 2: In black: posterior probability of $x = k$ for "around 20" from uniform priors on values x and interval radii y; in red: uniform prior on x; in blue: posterior for "between 10 and 30"

3.2.3 Main features

From this example, we see that the model captures the two main differences pointed out earlier concerning "around" and "between".

First of all, the vagueness of "around" is represented: semantically, the meaning of "around" does not specify a fixed interval. Moreover, in this example the posterior obtained for "around" captures sorites-susceptibility: we see that close values are assigned close posterior probabilities and that the posterior is a smooth function. For "between", by contrast, the meaning is crisp, and the posterior is given by a step function involving a "jolt" between some consecutive values (Smith 2008).

Secondly, peakedness is captured: starting from a flat prior, "around" outputs a nonuniform posterior in which values closer to the target number are more probable, whereas "between" outputs a flat posterior.

Finally, we note that in this example, the idea that the vagueness of "around" is a safeguard against error for the speaker is also present. From the listener's perspective, it is compatible with hearing "around 20" to assign a nonzero probability to the value 0, if indeed 0 was an option to her initially. Of course, the listener's posterior on 0 is very small and negligible in comparison to other values, but a speaker who would choose to specify a sharp interval using "between" incurs a higher risk of error in a situation in which they may suspect the listener to not rule out any value initially.[7]

[7] We return to this point, and to its philosophical significance, in section 10.

4 The ratio inequality

In the previous example, it was assumed that the priors were uniform, but this need not be the case in general. Suppose $P_L(x = k)$ is non-uniform to begin with. Then using "between i and j" will rescale that nonuniform prior to a new nonuniform posterior. Conversely, if values other than n are deemed initially more probable, the posterior may not show a peak on n for "around n". Despite that, we can still show that the effect of using "around", compared to using "between", is to increase the probability of 'central' values relative to less central ones, by assigning more weight to numbers nearer the target value.

This fact is captured by the following ratio inequality, which says that when a number k_1 is closer to the target value n than some other number k_2, then the ratio of the posterior probabilities of k_1 and k_2 upon hearing " x is around n" is larger than the ratio of their priors. More formally, let k_1 and k_2 be two integers such that $k_1 < k_2$, then:

$$\frac{P_L(x = n - k_1 \mid x \text{ is around } n)}{P_L(x = n - k_2 \mid x \text{ is around } n))} > \frac{P_L(x = n - k_1)}{P_L(x = n - k_2)} \tag{ix}$$

Let us show this. Call the first ratio (on the left-hand side of the above inequality) R_{post} and the second one R_{prior}. From Eq. (vi), it follows that:

$$\begin{aligned}
R_{\text{post}} &= \frac{P_L(x = n - k_1) \times \sum_{i \geq k_1} P(y = i)}{P_L(x = n - k_2) \times \sum_{i \geq k_2} P_L(y = i)} \\
&= \frac{P_L(x = n - k_1)}{P_L(x = n - k_2)} \times \frac{\sum_{i \geq k_1} P_L(y = i)}{\sum_{i \geq k_2} P_L(y = i)} \\
&= R_{\text{prior}} \times \frac{\sum_{y=k_1}^{y=k_2-1} P_L(y = i) + \sum_{y \geq k_2} P_L(y = i)}{\sum_{y \geq k_2} P_L(y = i)} \\
&= R_{\text{prior}} \times (1 + \frac{\sum_{y=k_1}^{y=k_2-1} P_L(y = i)}{\sum_{y \geq k_2} P_L(y = i)}) > R_{\text{prior}}
\end{aligned}$$

From the ratio inequality, it follows that the ratio of the posteriors for "around" is greater than the ratio of the posteriors for "between", since for "between", the ratio of the posteriors must be equal to the ratio of the priors. That is, let k_1 and k_2 belong to the interval $[i, j]$ specified by "between i and j", then:

$$\frac{P_L(x = k_1 \mid \text{between } i \text{ and } j)}{P_L(x = k_2 \mid \text{between } i \text{ and } j)} = \frac{P_L(x = k_1)}{P_L(x = k_2)}$$

As a result, the model predicts that even if the priors on possible values of x and on interval values of y are nonuniform to begin with, using "around n" instead of "between i and j" (for $i \leq n \leq j$) will give n a comparatively higher posterior. The ratio inequality matters, because it makes a prediction regarding the relationship between the posterior distributions generated by "between" adn "around" statements which is independent of the prior probability distribution. This prediction can in principle be investigated empirically.[8]

[8] We refer to Mortier (2019) for a report on some preliminary results.

5 A model of the speaker

Now that we have a precise model of the listener, we can provide a model for the speaker which will capture the idea that a speaker may prefer a vague 'around'-sentence over a precise 'between'-sentence if the 'around'-sentence leads the listener to a posterior probability distribution which is closer to that of the speaker.

In most approaches to pragmatics in formal semantics and philosophy of language, informativity is defined in terms of *entailment*, and information states are modeled as sets of possible worlds, i.e. propositions. Grice's maxim of quantity is then interpreted as a requirement that the speaker communicate the proposition that matches her information state. Consider a speaker who knows that the value of x is in a certain interval $[a, b]$, and does not know anything else. Then, if asked about the value of x, the best the speaker can do is to say something like "x is between a and b". In particular, an "around"-sentence fails to categorically rule out values that are outside of $[a, b]$, and so should never be preferred.

In our probabilistic setting, information states are more fine-grained, and are represented as probability distributions over states of the world ('worlds' for short). Imagine, in particular, that at some point the listener and the speaker had exactly the same information about some variable of interest (say the number of people who attended a certain party). Then the speaker makes a private observation that brings her to a new information state (a new probability distribution over the variable of interest). Let us assume that this private observation cannot be directly communicated. The goal of the speaker is then, in our setting, to find a message such that, when the listener will process it, the posterior probability distribution of the listener over the variable of interest will be as close as possible to hers.

What we need, therefore, is to specify the underlying *measure* that the speaker will use in order to assess how close to her own distribution the posterior distribution of the listener will be after processing her message. The measure we use comes from information theory, and is known as *Kullback-Leibler Divergence*, in the wake of recent decision-theoretic models of pragmatics (esp. models couched in the *Rational Speech Act* framework, see Frank et al. 2009; Goodman and Stuhlmüller 2013; Bergen et al. 2016). Before giving the formula for this measure, we first motivate it, and introduce the relevant information-theoretic background. Readers familiar with the concept of K-L divergence may go directly to section 5.2; others may find this a fruitful preamble, as the notion is often taken for granted.[9]

5.1 Information, surprisal, and Kullback-Leibler Divergence

Assume that the speaker and the listener start with the same prior probability distribution P over world-states, which we simply identify to the possible values of some variable of interest, noted x. Then the speaker makes a private observation o, as a result of which she has a new probability distribution over world-states, notated P_o. When the speaker uses a message m (for instance 'x is around 7', or 'x is between 5 and 9'), the listener processes it and ends up with a posterior distribution P_m (for instance, if m is 'x is around n', we have $P_m(x = k) = P(x = k|x \text{ is around } n)$, which we computed above).

Now, suppose after these two events happen, the state of the world, that is some number k, is observed by both the speaker and the listener. The more *unlikely* the observed number was

[9]For a general introduction to K-L divergence, see McElreath (2016, 179). The presentation we give is more specific to the communicative framework under discussion.

relative to their probability distributions, the more surprised they are, and the more information they get. In information theory, the information gained by an agent when observing k is equated to $-\log(P(x=k))$, where P is the probability distribution that represents the agent's epistemic state before the observation.

(6) a. Listener's surprisal when observing k, after having processed m:
$-\log(P_m(x=k))$

b. Speaker's surprisal when observing w, after having observed o:
$-\log(P_o(x=k))$

Since the listener may fail, after hearing the message, to fully recover the information that the speaker has, the listener whose probability distribution is P_m would be, on average, more surprised, when learning the true state of the world, than the speaker whose probability distribution is P_o (that is, observing the true state of the world world brings more information to someone who is not very knowledgeable than to someone who is more knowledgeable). Now, saying that the speaker wants to bring the listener to a state as close as possible to hers amounts, in this setting, to the idea that the speaker would like to minimize, across worlds, the difference in future surprisal between the listener and herself (ideally the listener would fully recover the information that the speaker has, and this average difference will be 0, in which case, if both observed the true state of the world, they would both be exactly as surprised).

(7) Difference of surprisals between Listener and Speaker after observing $x=k$:
$-\log(P_m(x=k)) - (-\log(P_o(x=k))) = \log(P_o(x=k)) - \log(P_m(x=k))$

The speaker does not know which world is in fact the case, so she wants to minimize the 'average', or 'expected' difference in surprisal between her and the listener in case they observed the actual world. But whose expectations should we use to compute this expected difference in surprisal? Suppose that there are only two worlds w_1 and w_2, and $P_o(w_1) = 0.9$ and $P_o(w_2) = 0.1$, while $P_m(w_1) = P_m(w_2) = 0.5$. Upon observing w_2, the speaker would be more surprised than the listener, so in this case the difference in surprisal between Listener and Speaker would be negative. The speaker has good reasons to think that w_2 is in fact very unlikely. It is much more likely for w_1 to be observed, and in this case the listener would be more surprised than the speaker. And this second possibility should receive more weight, since in fact, given the additional information that the speaker has, it is more likely to occur. That is, because the speaker's probability distribution results from a truthful observation, and therefore corresponds to a better epistemic state than that of the listener, the *expected* difference in surprisal between speaker and listener should be computed *from the perspective of the speaker*, and will be the following weighted average:

(8) $P_o(w_1) \times (\textit{difference in surprisal between Listener and Speaker if } w_1 \textit{ is observed})$
$+P_o(w_2) \times (\textit{difference in surprisal between Listener and Speaker if } w_2 \textit{ is observed})$

Generalizing from this simple case, we get the following:

(9) Expected difference in surprisal between Listener and Speaker, from the point of view of the

speaker who has observed o:

$$\sum_k P_o(x=k) \times [\log(P_o(x=k)) - \log(P_m(x=k))]$$
$$= \sum_k P_o(x=k) \times \log\left(\frac{P_o(x=k)}{P_m(x=k)}\right)$$

This quantity is known as the *Kullback-Leibler divergence* of P_m from P_o.[10]

(10) $D_{KL}(P_o||P_m) = \sum_k P_o(x=k) \times \log\left(\frac{P_o(x=k)}{P_m(x=k)}\right)$

5.2 The speaker's utility function and choice rule

The goal of a cooperative speaker who has observed o will be to pick a message m that *minimizes* the quantity we have just defined. To capture this idea, we can define a *utility* function which defines the *payoff* that the speaker gets from using message m if her information state if P_o, i.e. if she observed o, such that this payoff *increases* as $D_{KL}(P_o||P_m))$ decreases.[11]

(11) $U(m, o) = -D_{KL}(P_o||P_m)$

We now assume that, when making a choice between several messages $m_1, m_2, \ldots$, a speaker who has observed o picks the message m_i such that for every $j \neq i$, $U(m_i, o) > U(m_j, o)$.[12]

A case where the speaker prefers an 'around'-statement

We now provide a description of a case where a speaker would, according to our model, receive a higher payoff from using an "around"-statement than from using a "between statement". Our goal is to provide an existence proof, that is to show that our model makes it possible for an "around"-sentence to be a better message than any "between"-sentence.

We assume the value of interest x can range from 0 to 8, and that initially both the speaker and the listener have a uniform distribution over x (for any $k, P(k) = \frac{1}{9}$). Then the speaker makes an observation as a result of which she has a new probability distribution over worlds-states, notated

[10] For any two probability distributions P_1 and P_2, $D_{KL}(P_1||P_2)$ is always positive. This reflects the fact that it measures the gain in information when one starts with a distribution P_2 and makes an observation which results into a posterior distribution P_1. In case P_1 cannot be rationally reached from P_2 (because P_2 assigns probability 0 to some world-states that are assigned a non-null probability by P_1), the KL-divergence is infinite (see foonote 11).

[11] It is worth mentioning here that using this measure indirectly captures Grice's maxim of quality. This is for the following reason. Suppose that the speaker is in no position to exclude a certain world-state $x = j$, i.e. $P_o(x = j) > 0$. Suppose she picked a message that would in fact exclude this state, i.e. such that $P_m(x = j) = 0$. Then the quantity $log \frac{P_o(x=k)}{P_m(x=k)}$ can be viewed as infinite (because the denominator in the fraction is 0), and so one term of the sum in the formula above will be infinite, as a result of which $D_{KL}(P_o||P_m))$ is infinite, and $U(m, o)$ is infinitely negative.

[12] For simplicity, we ignore the possibility that two messages are exactly tied, i.e. have exactly the same utility. Furthermore, we assume here that the speaker is fully rational and picks the best message with probability 1. In Rational Speech Act models, the speaker is typically not assumed to be fully rational. Rather, she follows a so-called 'SoftMax'-rule whereby she is more likely to use the best message than the second best, more likely to use the second best than the third best, etc., but nevertheless does not pick the best message with probability 1. This difference is not important at this point. However in section 6, where we develop a fully explicit RSA model, we use the SoftMax rule.

P_o, which categorically excludes only the values 0 and 8, but is extremely biased towards central values, giving a 96% probability to the interval $[3, 5]$ (cf. Table 1).

Table 1: Hypothetical Speaker's distribution P_o.

k	0	1	2	3	4	5	6	7	8
$P_o(k)$	0	0.01	0.01	0.16	0.64	0.16	0.01	0.01	0

It is clear that the optimal 'between'-message for the speaker is 'x is between 1 and 7', since this message exactly specifies the support of the speaker's distribution. But the speaker could also use the message 'x is around 4'. Now, we consider the posterior probability distributions of the listener after processing both messages (Table 2), assuming the listener has a uniform prior on the intervals expressed by "around". After processing the 'between'-message, the listener ends up with a uniform distribution on the interval $[1, 7]$. After the 'around'-message, using Eq. (viii) from Section 3.2.2, the listener ends up with a probability distribution that does not categorically rule out any value, but gives more weight to central values.

Table 2: Listener's posterior distributions after hearing 'between 1 and 7' and 'around 4'.

k	0	1	2	3	4	5	6	7	8
$P_{between}(k)$	0	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0
$P_{around}(k)$	0.04	0.08	0.12	0.16	0.20	0.16	0.12	0.08	0.04

While neither of these distributions is intuitively close to the hypothesized speaker's distribution, the one that results from the *around*-message is biased towards central values, like the hypothesized speaker distribution. On the other hand, it fails to exclude the values 0 and 1, which the speaker's distribution excludes. When we now compute the KL-divergence of each of these distributions from the speaker's distribution, it turns out that we get a smaller value with $P_{around}(k)$ than with $P_{between}(k)$, hence the "around"-sentence is better at reducing the distance between the listener's distribution and the speaker's distribution.[13] We have $D_{KL}(P_o||P_{between}) = 0.89$, while $D_{KL}(P_o||P_{around}) = 0.65$. This translates into the following speaker utilities for each message: U('x is around 4', o) = -0.65 and U('x is between 1 and 7', o) = -0.89. The speaker will thus receive a higher utility from the "around"-message, and is therefore predicted to use it.

6 A full RSA-model for 'around'

So far, our model of the listener does not take into account the fact that the speaker chooses her message strategically (as just discussed), and that the listener can use her knowledge of the speaker's choice rule to derive additional inferences. Rather, our listener, when interpreting an "around n"-message, simply conditionalizes her joint probability distribution on x (the variable of interest) and y (which determines the length of the interval corresponding to an "around"-statement) with the

[13] The reason this is the case is that for every value k in $[1, 7]$, $P_{around}(k)$ is closer to $P_o(k)$ than $P_{between}(k)$ is. $P_{between}$ wins only for the extreme values 0 and 8 but since these values have anyway a null probability of occurrence according to P_o, they do not play any role in the computation of the expected difference in surprisal, which is computed from the point of view of the speaker.

information that x is in the interval $[n - y, n + y]$, but does not derive any inference about the speaker's epistemic state.

Now, a *pragmatic* listener might reason that if the speaker picked an "around"-sentence rather than a "between"-sentence, this might precisely be because the speaker's probability distribution is biased towards central values, so that the "around"-sentence was a better sentence to use than one based on "between". Such an extra inference might then strengthen the conclusion that central values are more likely than peripheral values. And an even more sophisticated speaker might take this into account when choosing her message, making the "around"-message even more appropriate when the speaker's epistemic state is biased towards central values.

In this section, we provide a model which can capture these extra inferences. This model is a particular version of the Rational Speech Act framework. It first defines a listener of level-0 who is just the one we have defined in section 3.2. This listener simply conditionalizes her joint distribution on x and y with the information carried by an "around"-sentence in virtue of its linguistic meaning, namely the proposition that x is the interval $[n - y, n + y]$, where both x and y are variables whose values are not known.[14]. Importantly, this basic listener draws no inference about the epistemic state of the speaker. Then a first-level pragmatic speaker is defined along the lines of the speaker model introduced in section 5. But then we can define a pragmatic listener (called the 'first-level pragmatic listener', L^1 for short) who knows that she received a message from the first-level pragmatic speaker, and who updates (using Bayes' rule) her probability distribution over both the variable of interest x, and a variable o ranging over the possible epistemic states of the speaker. On this basis, we can then define a second-level pragmatic speaker who chooses her message strategically, still with the goal of making the listener's epistemic state about x (the variable of interest) as close as possible to hers, based on the assumption that the listener she is speaking to is the first-level pragmatic listener. This process can continue indefinitely, and defines an infinite sequence of listeners and speakers – the higher we are in the sequence, the more pragmatically sophisticated the speaker and the hearer are.

We now proceed to describe such a model in detail. We first describe it 'in the abstract' and then present a specific implementation. Our goal is to show how the basic effect we have been discussing can get amplified through pragmatic reasoning.

6.1 Set-up

We assume that the listener cares about the value of some variable x which ranges over the natural numbers between 0 and 8. Before the speaker makes a private observation, they share a joint probability distribution P over pairs $\langle x, o \rangle$, where x is the variable of interest, and o ranges over a set of *observations* O that the speaker could in principle make (we will specify the set of observations as well as other important ingredients of the model in Section 7). Then the speaker makes a private observation o_j, as result of which her new probability distribution is P_{o_j}, defined by $P_{o_j}(x = k) = P(x = k \mid o = o_j)$. Furthermore, the listener has a prior probability distribution over the variable y, which determines the interpretation of *around*, as discussed above. We assume that the prior distribution on y, x and o is such that y is probabilistically independent of x and o (x and o are

[14]Importantly, as noted in footnote 4, we substantially depart from RSA models with lexical uncertainty (e.g., Lassiter and Goodman 2017), which we discuss in section 8 and Appendix A. In such models, the literal listener is relativized to a fixed value for y, and there are as many literal listeners as they are possibly values for y. It is only at the level of the first *pragmatic* listener that uncertainty about the value of y is factored in.

not independent, since the observation that one is likely to make will typically depend in part on the value of x).

The set M of possible messages is the following: *Between 0 and 8*, *between 1 and 7*, *between 2 and 6*, *between 3 and 5*, *exactly 4*, *Around 4*.[15]

The literal listener L^0 is characterized by the following update rule, which defines $L^0(x = k, o = o_i|m)$, the distribution over $\langle x, o\rangle$ which characterizes the level-0 listener after she has processed the message m:[16]

(12) a. If m is of the form *between a and b* (treating "exactly 4" as equivalent to "between 4 and 4"), then:
$L^0(x = k, o = o_j|m) = 0$ if k is not in the interval $[a, b]$;
otherwise, $L^0(x = k, o = o_j|m) \propto P(x = k, o = o_j)$.[17]

b. If m is the message 'around 4' then:
$L^0(x = k, o = o_j|m) \propto P(x = k, o = o_j) \times \sum_{i\geq|4-k|} P(y = i)$

6.2 The level-1 pragmatic speaker

The level-1 pragmatic speaker believes she talks to L^0. If she made the observation o_j, she wants to use a message m such that the posterior distribution of L^0 over x after processing message m is maximally close to her own epistemic state, namely P_{o_j}. Now, this means she does not care about the listeners's beliefs about o, but only about the listener's beliefs about x. Let us note L^0_m the distribution over x of the literal listener after she has processed m (L^0_m is not a joint distribution over

[15] This is of course a gross oversimplification, since we only consider messages which are 'centered' on 4. The only reason we do this is that this limitation makes the model reasonably tractable and intelligible. Given that the set of *observations* we consider in section 7.1 will result in posterior distributions which are themselves centered on 4, it is likely that other "between"-statements that would not be centered on 4 would be generally suboptimal for the speaker compared to the messages that we include in the model.

[16] Regarding (12-b), note that in contrast with Equation (vi), we use here the joint distribution $P(x, o)$ instead of just $P(x)$. We obtain this formula in the same way as we obtained the one in Eq. (vi). Bayes'rule gives us:

$$L^0(x = k, o = o_j, y = i \mid d(x, 4) \leq y) \propto P(d(x, 4) \leq y \mid x = k, o = o_j, y = i) \times P(x = k, o = o_j).$$

Now, obviously, $P(d(x, 4) \leq y \mid x = k, o = o_j, y = i)$ is either 0 or 1, depending on whether $d(k, 4) \leq y$, and the value of o does not play any role (that is, the literal meaning of the message does not carry any direct information about o, the observation that the speaker made, but only about x). So the above equation simplifies to:

$$L^0(x = k, o = o_j, y = i \mid d(x, 4) \leq y) \propto P(d(x, 4) \leq y \mid x = k, y = i) \times P(x = k, o = o_j),$$

and the rest of the computation proceeds in the same way as in Equation (vi).

Note that we have:

$L^0(x = k|m) \propto \sum_{o_h\in O} L^0(x = k, o = o_h|m)$

$= \sum_{o_h\in O} [P(x = k, o = o_h) \times \sum_{i\geq|n-k|} P(y = i)]$

$= \sum_{i\geq|n-k|} P(y = i) \times \sum_{o_h\in O} P(x = k, o = o_h)$

$= \sum_{i\geq|n-k|} P(y = i) \times P(x = k) = P(x = k) \times \sum_{i\geq|n-k|} P(y = i)$.

Hence we recover Equation (vi).

[17] L^0, when processing a "between"-message, updates her distribution by assigning a probability 0 to values that are incompatible with the literal meaning of the message, and multiply the probabilities of the remaining values by a constant term so that they sum up to 1.

x and o, it is a distribution over x alone that results from marginalizing the conditional distribution $L^0(x = \ldots, o = \ldots \mid m)$ over o): $L^0_m(x = k) = L^0(x = k \mid m) = \sum\limits_{o_h \in O} L^0(x = k, o = o_h|m)$

Based on the discussion in Section 5, the utility function U^1 of the level-1 pragmatic speaker is given by:[18]

(13) $\quad U^1(m, o_j) = -D_{KL}(P_{o_j}||L^0_m)$

Now, as is standard in RSA models, we do not assume a fully rational speaker. Rather, the speaker is only approximately rational, and uses a so-called *SoftMax* rule. The higher the utility of a message is, the more likely is the speaker to use it, but the best message is not used with probability 1. The SoftMax rule is parametrized by a 'temperature'-parameter λ, a positive real number. The higher λ is, the more rational the speaker is, meaning that the probability of using the best message approaches 1 as λ increases. The speaker S^1 is defined by the conditional probability of using a message given that a certain observation was made, and the SoftMax function is used to define this probability:[19]

(14) $\quad S^1(m \mid o_j) \propto \exp(\lambda \times U^1(m, o_j))$

6.3 Higher-order listeners and speakers

Of special interest to us is the first pragmatic listener L^1, who interprets messages under the assumption that the author of the message is S^1. L^1 simply applies Bayes rule, and makes a *joint inference* about both x and o. Note that at this point no further inference takes place about y (which enters into the interpretation of "around" for L^0). The listener L^1 uses the same prior distribution P over x and o as L^0 (this distribution basically characterizes what was common knowledge between speaker and addressee before the speaker made any observation).

Bayes' rule gives us:

$$L^1(x = k, o = o_j \mid m) \propto P(x = k, o = o_j) \times S^1(m \mid x = k, o = o_j) \qquad \text{(x)}$$

Now, note that the speaker's behavior only depends on her observation — it depends on the value of x only to the extent that the observations she is likely to make are not the same across different values of x. The speaker does not directly observe the value of x, but only receives information through the observation she made, and she decides which message to use on the basis of this observation (not on the value of x, which she typically does not know — all the knowledge she has about x is contained in her distribution $P_{o=o_j}$). This means that $S^1(m \mid x = k, o = o_j) = S^1(m \mid o = o_j)$. So we have:

$$L^1(x = k, o = o_j \mid m) \propto P(x = k, o = o_j) \times S^1(m \mid o = o_j) \qquad \text{(xi)}$$

[18] Our utility function is actually slightly different from the one used in most RSA models, where the speaker is modeled as wanting to make the hearer correctly identify her epistemic state on top of the true state of the world. We discuss this point in more detail in Section 8.

[19] Dispensing with the proportionality notation $\propto$, we can write, more explicitly:
$S^1(m_i \mid o_j) = \dfrac{\exp(\lambda \times U^1(m, o_j))}{\sum\limits_{m_i \in M} \exp(\lambda \times U^1(m_i, o_j))}$. When λ tends to infinity, this quantity tends to 1 if m_i is the message that receives the highest utility.

We notate L^1_m the probability distribution over x for the listener L^1 of m (L^1_m is the marginal distribution over x for a listener who receives message m, not a joint distribution over x and o). We have:

(15) $L^1_m(x = k) = L^1(x = k \mid m) = \sum_{o_h \in O} L^1(x = k, o = o_h \mid m)$
$\propto \sum_{o_h \in O} P(x = k, o = o_h) \times S^1(m \mid o = o_h)$

We can then generalize this logic and define higher-level speakers and listeners as follows:

(16) For $n \geq 1$,
a. $U^{n+1}(m, o_j) = -D_{KL}(P_{o_j} || L^n_m)$
b. $S^{n+1}(m \mid o_j) \propto \exp(\lambda \times U^{n+1}(m, o_j))$

(17) For $n \geq 1$, $L^n_m(x = k) \propto \sum_{o_h \in O} P(x = k, o = o_h) \times S^n(m \mid o_h)$

As we shall see in the next section, the effect of "around" tends to be magnified for higher-level listeners and speakers compared to what happens at the level of the literal listener, in that the bias towards central values is increased.

7 A concrete implementation of the interactive model

We now provide a concrete implementation of the model we have just described. The goal here is not to propose a realistic model — as we shall see, some choices will be quite arbitrary — but to illustrate the point that such an explicit model predicts that the basic effect we have been discussing (the fact that "around"-statements can be used to convey the shape of a probability distriubtion) can be amplified by pragmatic reasoning.

7.1 Observations and prior probability distributions

As we saw, we assume that the variable of interest x ranges over the natural numbers between 0 and 8. We consider that nine observations o are possible. From the point of view of the model, the only thing that matters is the effect of an observation o_j on an observer who starts with a prior probability distribution P over x, i.e. we need to specify P_{o_j}, i.e. the different epistemic states the speaker could be in after having made an observation. Here we are interested in comparing the speaker's behavior in epistemic states which have the same support but a different 'shape' (from uniform distributions to distributions biased towards central values). The nine observations we consider are characterized by Table 3, where each column is the probability distribution induced by a specific observation, and each line is a possible value for x.

Table 3: Posterior distributions resulting from observations ($P(x|o)$)

x \ o	=4	u_3_5	u_2_6	u_1_7	u_0_8	p_3_5	p_2_6	p_1_7	p_0_8
0	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.14	0.11	0.00	0.00	0.02	0.03
2	0.00	0.00	0.20	0.14	0.11	0.00	0.06	0.09	0.11
3	0.00	0.33	0.20	0.14	0.11	0.25	0.25	0.23	0.22
4	1.00	0.33	0.20	0.14	0.11	0.50	0.38	0.31	0.27
5	0.00	0.33	0.20	0.14	0.11	0.25	0.25	0.23	0.22
6	0.00	0.00	0.20	0.14	0.11	0.00	0.06	0.09	0.11
7	0.00	0.00	0.00	0.14	0.11	0.00	0.00	0.02	0.03
8	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00

As we can see, the first five columns correspond to five observations that give rise to a uniform distribution on an interval centered on 4 (from the observation "=4", where the speaker can conclude from her observation that $x = 4$, to the observation "u_0_8", which results in a uniform distribution on the full range of x). The last four observations correspond to distributions with supports $[3, 5]$, $[2, 6]$, $[1, 7]$ and $[0, 8]$, respectively, with a bias towards values closer to 4. These were obtained from binomial distributions with parameter 0.5, shifted to the relevant intervals. An observation named *u_a_b* corresponds to a posterior distribution which is uniform on its support $[a, b]$, while *p_a_b* corresponds to a posterior distribution with support $[a, b]$ which is 'peaked', in the sense of being biased in favor of more central values.

To build the joint distribution on $\langle x, o\rangle$, we will assign first probabilities to each observation, and derive the joint distribution from the fact that

$$P(x = k, o = o_i) = P_{o_i}(x = k) \times P(o_i). \tag{xii}$$

The resulting marginal probability distribution over x will then be given by

$$P(x = k) = \sum_{o_i \in O} P(o_i) \times P_{o_i}(x = k). \tag{xiii}$$

Now, values close to 4 belong to the support of more distributions than values further from 4, and all the distributions induced by an observation are either uniform or biased towards values closer to 4. As a result of this setup, the marginal distribution over x will itself be non-uniform, and biased towards central values.[20] We chose to assign higher probabilities to observations that yield distributions with a larger support, so as to not penalize too much peripheral values. Specifically, we assigned the following weights to each observation, which we then normalized to get a probability distribution:

[20]This is by no means a necessary choice. But to make an already complex model reasonably tractable, we chose to restrict the set of possible observations to those that are 'centered' on 4, as this suffices to make our main point.

Table 4: Probabilities over observations ($P(o = o_i)$)

Observation	=4	u_3_5	u_2_6	u_1_7	u_0_8	p_3_5	p_2_6	p_1_7	p_0_8
Non-normalized Weight	1	4	16	64	256	1	4	16	64
Probability	$\frac{1}{426}$	$\frac{2}{213}$	$\frac{8}{213}$	$\frac{32}{213}$	$\frac{128}{213}$	$\frac{1}{426}$	$\frac{2}{213}$	$\frac{8}{213}$	$\frac{32}{213}$

The resulting marginal probability distribution on x (given by Eq. (xiii)) is the following:

Table 5: Prior Probability Distribution over x

x	0	1	2	3	4	5	6	7	8
$P(x)$	0.07	0.09	0.12	0.14	0.16	0.14	0.12	0.09	0.07

While the most likely observation, namely u_0_8, is the one that would lead to a posterior distribution that is uniform on $[0, 8]$ (cf. Table 3), the prior distribution on x is not itself uniform (cf. Table 5) but biased towards central values. That is, this observation leads the observer to raise the probability of values that are far from 4, and to decrease that of central values.

Finally, the joint probability distribution on $\langle x, o \rangle$, given by Eq. (xii), can be described by the following table:

Table 6: Joint Prior Distribution on x and o ($P(x, o)$)

x \ o	=4	u_3_5	u_2_6	u_1_7	u_0_8	p_3_5	p_2_6	p_1_7	p_0_8
0	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.001
1	0.000	0.000	0.000	0.021	0.067	0.000	0.000	0.001	0.005
2	0.000	0.000	0.008	0.021	0.067	0.000	0.001	0.004	0.016
3	0.000	0.003	0.008	0.021	0.067	0.001	0.002	0.009	0.033
4	0.002	0.003	0.008	0.021	0.067	0.001	0.004	0.012	0.041
5	0.000	0.003	0.008	0.021	0.067	0.001	0.002	0.009	0.033
6	0.000	0.000	0.008	0.021	0.067	0.000	0.001	0.004	0.016
7	0.000	0.000	0.000	0.021	0.067	0.000	0.000	0.001	0.005
8	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.001

7.2 Other parameter of the models

The variable y, which determines the size of the interval denoted by "around 4" and enters into the interpretation of "around 4" at the level of the literal listener, ranges from 0 to 4 (recall that $y = i$ means that the intended interval for "around 4" is [4-i, 4+i]). We assume that the prior probability distribution over y, which is used by the literal listener, is uniform.

As discussed in section 6.1, we are assuming that the speaker has six messages at her disposal (*Exactly 4*, *between 3 and 5*, *betwen 2 and 6*, *between 1 and 7*, *between 0 and 8*, *around 4*). We set the 'temperature parameter' which enters into the utility function of the speaker to 10.

7.3 Predictions of the model

First we look at the literal listener L^0, in Table 7. Each cell of the table represents the probability assigned by L^0 to a number 0 and 8 after having interpreted a given message ("b. 3 and 5" represents "between 3 and 5").

Table 7: Probabilities assigned by L^0 to each value for x after processing a message m ($L^0(x \mid message)$)

x \ Message	Exactly 4	b. 3 and 5	b. 2 and 6	b. 1 and 7	b. 0 and 8	around 4
0	0.00	0.00	0.00	0.00	0.07	0.02
1	0.00	0.00	0.00	0.11	0.09	0.06
2	0.00	0.00	0.17	0.13	0.12	0.11
3	0.00	0.32	0.21	0.17	0.14	0.18
4	1.00	0.36	0.23	0.18	0.16	0.25
5	0.00	0.32	0.21	0.17	0.14	0.18
6	0.00	0.00	0.17	0.13	0.12	0.11
7	0.00	0.00	0.00	0.11	0.09	0.06
8	0.00	0.00	0.00	0.00	0.07	0.02

Note that the message *between 0 and 8* is in fact completely uninformative. As expected, then, the posterior distribution that results from this message is identical to the prior distribution. All other "between"-messages assign 0 to the numbers they exclude. Finally, the "around"-message does not assign 0 to any number, but results in a distribution that is much more biased in favour of central values than the prior distribution was, as expected (one can compare it with the column for *between 0 and 8*, which corresponds, as noted, to the prior distribution). Note also that it assigns a higher value to 4 than the posterior distribution obtained after processing *between 2 and 6*, *between 1 and 7*, and *between 0 and 8*, following the logic discussed in 4. So it might end up being the best message to use for a speaker who is uncertain but assigns a high probability to central values.

We now turn to the behavior of the level-1 speaker, who talks to this listener. The following table represents the probability of using a certain message depending on the observation that the speaker made.

Table 8: S^1's probability of choosing a message depending on the observation made ($S^1(m \mid o)$)

Observation \ Message	Exactly 4	b. 3 and 5	b. 2 and 6	b. 1 and 7	b. 0 and 8	around 4
=4	1.00	0.00	0.00	0.00	0.00	0.00
u_3_5	0.00	0.98	0.01	0.00	0.00	0.01
u_2_6	0.00	0.00	0.82	0.07	0.02	0.09
u_1_7	0.00	0.00	0.00	0.69	0.16	0.15
u_0_8	0.00	0.00	0.00	0.00	0.93	0.07
p_3_5	0.00	0.97	0.01	0.00	0.00	0.01
p_2_6	0.00	0.00	0.68	0.06	0.01	0.25
p_1_7	0.00	0.00	0.00	0.27	0.06	**0.66**
p_0_8	0.00	0.00	0.00	0.00	0.14	**0.86**

Now, when the speaker's distribution is uniform across a certain interval (first five lines), the speaker has a very high probability of using the corresponding *between*-statement. If the speaker's distribution has support $[0, 8]$ but is biased towards central values (last line), the speaker prefers the *around*-statement. Importantly, even when the speaker is able to categorically exclude 0 and 8 and has a distribution which is biased in favor of central values (*p_1_7*), she prefers to use the *around*-statement rather than the corresponding *between*-statement, following the logic of what we discussed in Section 5.2. She has also a non-insignificant probability of using the *around*-statement when her distribution is a peaked distribution with support is $[2, 6]$

Now, we consider the 1st-level pragmatic listener (Table 9). This listener knows that the speaker's choice of message is governed by Table 8. So when hearing the *around*-statement, she will infer that the speaker is most likely in the epistemic state that results from the observations *p_0_8* or *p_1_7*. She will update her distribution over x on this basis.

Table 9: Probabilities assigned by L^1 to each value for x after processing a message m ($L^1(x \mid message)$)

x \ Message	Exactly 4	b. 3 and 5	b. 2 and 6	b. 1 and 7	b. 0 and 8	around 4
0	0.00	0.00	0.00	0.00	0.10	0.02
1	0.00	0.00	0.00	0.13	0.11	0.05
2	0.00	0.00	0.18	0.14	0.11	0.11
3	0.00	0.32	0.21	0.15	0.12	0.19
4	1.00	0.37	0.23	0.16	0.12	0.24
5	0.00	0.32	0.21	0.15	0.12	0.19
6	0.00	0.00	0.18	0.14	0.11	0.11
7	0.00	0.00	0.00	0.13	0.11	0.05
8	0.00	0.00	0.00	0.00	0.10	0.02

While the posterior distribution of L^1 (cf. Table 9) after processing *around 4* is slightly less peaked than it was for L^0, the posterior distributions induced by the messages *between 0 and 8*, *between 1 and 7* and *between 2 and 6* are themselves significantly flatter (more uniform) than they were for L^0, so the contrast in interpretation between statements based on *between* and the one based on *around* is maintained (and even amplified if we compare the ratios, across distributions, between central values and peripheral values that have a non-null probability).

We can then look at even higher-order listeners and speakers. After a few iterations, we reach a near-steady state where further iterations do not change anything significantly. Tables 10 and 11 display the behavior of the level-5 speaker and listener, respectively.

Table 10: S^5's probability of choosing a message depending on the observation made ($S^5(m \mid o)$)

Observation \ Message	Exactly 4	b. 3 and 5	b. 2 and 6	b. 1 and 7	b. 0 and 8	around 4
=4	1.00	0.00	0.00	0.00	0.00	0.00
u_3_5	0.00	0.96	0.01	0.00	0.00	0.03
u_2_6	0.00	0.00	0.78	0.03	0.00	0.19
u_1_7	0.00	0.00	0.00	0.87	0.08	0.06
u_0_8	0.00	0.00	0.00	0.00	1.00	0.00
p_3_5	0.00	0.96	0.01	0.00	0.00	0.03
p_2_6	0.00	0.00	0.41	0.01	0.00	**0.57**
p_1_7	0.00	0.00	0.00	0.05	0.00	**0.94**
p_0_8	0.00	0.00	0.00	0.00	0.01	**0.99**

Table 11: Probabilities assigned by L^5 to each value for x after processing a message m ($L^5(x \mid message)$)

x \ Message	Exactly 4	b. 3 and 5	b. 2 and 6	b. 1 and 7	b. 0 and 8	around 4
0	0.00	0.00	0.00	0.00	0.11	0.00
1	0.00	0.00	0.00	0.14	0.11	0.03
2	0.00	0.00	0.18	0.14	0.11	0.11
3	0.00	0.32	0.21	0.15	0.11	0.22
4	1.00	0.37	0.22	0.15	0.11	0.27
5	0.00	0.32	0.21	0.15	0.11	0.22
6	0.00	0.00	0.18	0.14	0.11	0.11
7	0.00	0.00	0.00	0.14	0.11	0.03
8	0.00	0.00	0.00	0.00	0.11	0.00

On the speaker's side (Table 10), we see that the speaker will prefer the around message when her observation leads to a peaked distribution on the intervals $[2, 6]$, $[1, 7]$ and $[0, 8]$, and that when her distribution is uniform, she goes for the *between*-message that corresponds to the support of her distribution. On the listener's side (Table 11), we end up with distributions that are quite flat for all *between*-messages, and biased towards central values for the *around*-sentence. The recursive aspect of the model led to an amplification of the basic phenomenon observed at the level of the literal Listener L^0 and the first-level pragmatic speaker S^1. In particular, the level-5 speaker now uses the "around"-statement also when her distribution is peaked with support $[2, 6]$.

8 Comparison with the Lexical Uncertainty approach

Our full model is couched in the RSA framework for pragmatics. In the RSA literature, the most prominent model for dealing with semantic underspecification (e.g., gradable adjectives) is the so-called Lexical Uncertainty model (LU model for short, cf. Bergen et al. 2016; Lassiter and Goodman 2017). However, that literature does not discuss one of the main points of our paper: the fact that vague language might allow a speaker who is not fully informed about some topic under discussion to communicate information about the *shape* of her probability distribution. As it turns out, we can show that without fundamental modifications, the LU model is unable to reproduce the qualitative predictions of our model, especially regarding the speaker's behavior.

Concerning speaker uncertainty, Lassiter and Goodman (2017) simply do not incorporate in their model the possibility that the speaker is not fully informed about the value of interest (say, someone's height, in relation to the use of *tall*). They consider a speaker who knows, say, Mary's height, and can use messages such as *Mary is tall*, *Mary is not tall*, *Mary is short*, *Mary is not short*. The goal of the model is to predict the *interpretation* of such messages on the listener's side, and the listener is assumed to reason under the assumption that the speaker is fully informed. Their paper focuses on how, when processing such a message, the first-level pragmatic listener L^1 updates their probability distribution over Mary's height.

Even though this is not done in Lassiter and Goodman (2017), the LU model can of course include the possibility of a non-fully-informed speaker (see, e.g, Bergen et al. 2016, in connection with so-called Hurford Disjunctions). It is straightforward to apply the LU model to "around" in the general case where the speaker is not fully informed, by treating the size of the intended interval in the same way as the threshold for gradable adjectives is treated in LU models. Since such a model defines the utility function of the speaker in terms of Kullback-Leibler divergence,[21] one might think — and this was our initial expectation — that it would reproduce the qualitative predictions of our model, especially the prediction that the probability of using an"around"-sentence might depend on the *shape* of the speaker's distribution, rather than just its support. This is in fact not the case, for a fundamental reason, namely the following mathematical fact, proven in Appendix A:

(18) Let two observations o_1 and o_2 be such that the supports of the conditional distributions $P(x = k|o_1)$ and $P(x = k|o_2)$ are identical (that is $P(x = k|o_1) > 0$ iff $P(x = k|o_2) > 0$). In the LU model, for every message m, at every step of the recursion, we have $S(m|o_1) = S(m|o_2)$.

This means that the speaker's choice of a message only depends on the *support* of her distribution, not on its shape, and so one core idea of our own proposal cannot be captured in the LU model. This is not to say that the LU model predicts none of the effects we discuss. The first-level listener, in the LU model, does end up with a posterior distribution that favors central values after processing an "around"-statement (though, in our simulation, to a much lower extent than in our model). However, the speaker does not take this fact into account when choosing her

[21] Technically speaking, many RSA models are written without explicit reference to Kullback-Leibler divergence, but it is easy to show, that they are fully equivalent to models that use the Kullback-Leibler divergence in the utility function of the speaker.

message, and as a result this effect on the listener tends to fade away when we move higher up in the recursive sequence of listeners (because in contrast with our model, pragmatic listeners can only draw inferences about the support of the speaker's distribution, not its shape).

To be more precise, the LU model with a non-fully informed speaker differs from our own in two main respects.

First, semantic underspecification (in our case, the size of the interval for *around*) is dealt with differently. In the LU model, applied to *around*, the literal listener L^0 is relativized to a specific interpretation function, and would interpret an "around n"-statement as meaning "between a and b", where a and b are set by the interpretation function. So there are as many literal listeners as there are interpretation functions (i.e. ways of interpreting "around", since the interpretation function does not matter for other messages). Likewise, the first-level speaker is relativized to an interpretation function, and chooses her message on the assumption that the listener she is talking to is a literal listener relativized to the same interpretation function. So at the level of the literal listener and the first-level speaker, nothing interesting happens for an "around"-sentence, which is just treated in the same way as a "between"-statement. It is at the level of the first-level pragmatic listener (and similarly for speaker) that reasoning about the interpretation function (i.e. the intended interval for *around*) takes place, in the sense that this first-level pragmatic listener is no longer relativized to a specific interpretation function, but reasons probabilistically about the interpretation function, treated as a random variable. In our model, the listener of level 0 is right away interpreting "around" by taking into account its multiplicity of more precise interpretations.

The second difference is that in the LU model (as in most extant RSA models) the utility function of a non fully informed speaker is subtly different from ours (as hinted in footnote 18). It is defined in terms of the Kullback-Leibler divergence of the listener's joint distribution on $\langle x, o \rangle$ from the speaker's distribution on $\langle x, o \rangle$. In contrast, our speaker wants to minimize the Kullback-Leibler divergence of the listener's distribution over x from the speaker's distribution over x. Hence, the utility function used in most RSA-models views the speaker as caring not only about bringing the listener's distribution over the variable of interest as close as possible to hers, but also about communicating to the listener her private epistemic state about this variable. In contrast, the utility function in our model views the speaker as wanting to bring the listener's distribution over the variable of interest as close to hers, but not as caring about whether the listener correctly identified her epistemic state.

In many RSA models, the choice between these two options does not greatly affect qualitative predictions. And indeed, from the standpoint of our model, simulations show that if we use the utility function standardly used in RSA models (including the LU model), we can still reproduce the qualitative predictions of our model (cf. model described in Appendix C.1). From the standpoint of the LU model, however, it turns out that this choice is highly consequential. One can construct (cf. Appendix C.2) a version of the LU model where the utility function is defined as in our model in terms of the KL-divergence of the listener's and speaker's distributions over x (rather than their joint distributions over $\langle x, o \rangle$), but in which the result in (18) no longer holds: simulations show that, at least for very high values of the temperature parameter, the amended model can make predictions which are qualitatively similar to ours.

The upshot of this discussion is that it is the combination of a specific architectural choice (postponing to L^1 the listener's reasoning about the size of the interval for "around") and the choice of a specific utility function that makes the LU model unable to predict that the shape of the speaker's distribution, and not just its support, plays a role in the speaker's choice of a message.

To forestall the limitation we state in (18), one can either drop the architectural feature (e.g., by moving to a model like ours where the literal listener already treats the interval size for *around* as a random variable), or change the utility function along the lines of our model. Our discussion thus provides an argument for potential amendments to the most standard treatment of semantic underspecification in RSA models.[22]

9 Further issues

9.1 Rounding

While our model predicts the contrasts highlighted in section 2 regarding the use of "around" and "between", the central assumption we made of a speaker who is uncertain about the state of the world sets aside further facts concerning the use of "around".

The first of those concerns rounding, namely cases in which the speaker is perfectly informed about the numerical value of interest, but may nevertheless choose to use "around" instead of reporting the exact value. Consider a teacher who knows that 19 children attended her class. When asked "how many children did you have in class today?" she may respond by uttering: "around 20 children".[23] In that case, it would be incorrect for the listener to infer that 20 is the most likely observation made by the speaker, since by assumption 19 is the most likely value. This appears to contradict our model.

However, this is a case in which the speaker is using a round number to inform the listener. The effect produced would be very different, and the utterance may even sound odd, if the speaker responded: "around 16 children". For the latter, the listener ought to make the inference that the speaker is not perfectly informed about the state of the world, because 16 is *not* a round number and reports the number with a precision of 1 unit, and 1 is the finest possible granularity in this case (contrast with "around 16 kilometers" to report a distance run by bike, which may be used to round off a distance expressible with finer granularity in hundreds of meter or in meters).[24]

There is, therefore, a clear interaction between the use of "around" and considerations of granularity. This interaction is broadly consistent with our main prediction on the listener's side. By saying "around 20 children", the speaker still conveys that values closer to 20 (with the exception of 20 itself) are more likely than either 10 or 30. In that sense one prediction of our model of the interpretation of "around" appears to survive: when n is a round number, "around n" conveys that values closer to n (but not n itself if the speaker can be thought to be perfectly informed) are more likely than more remote values. That being said, our model does not involve any consideration of granularity, and excludes the possibility that a fully informed speaker will use an "around"-sentence with a significant probability. We leave a formal treatment of this interaction between roundness,

[22] As discussed in Bergen et al. (2016), though, the LU model's treatments of Hurford Disjunctions and Manner Implicatures both rely on the architecture of the LU model, in the sense that their predictions cannot be reproduced in models where reasoning about semantic underspecification takes place at L^0 or S^1, rather than L^1.

[23] "Approximately 20" may be more natural to use than "around 20" in case the speaker is perfectly informed and does rule out 20; however, we think it is possible to use "around" in the sense of "approximately". There are obviously subtle differences in meaning between"around", "about", and "approximately", which we set aside in this paper.

[24] Reporting "around 16 children" is odd but not ruled out, for instance if the speaker tries to remember how many children attended class, by remember how many children sat in each row, adds up the numbers, and wants to convey that the value obtained may be inaccurate.

granularity and the use of "around" by a perfectly informed speaker to another occasion.[25]

9.2 Common priors

Another limitation of the model concerns the common prior assumption. This limitation is not specific to our approach, it is shared by RSA approaches and by most game-theoretical set-ups. It is needed for the recursive definition of listeners and speakers to make sense from a normative point of view. It assumes that the interlocutors have access to their own probability distribution, but also that the speaker and the listener shares the same prior probability distribution over the variable of interest before the speaker makes a private observation about the state of the world. Those assumptions are obviously disputable, as they are most of the time violated in real life. Although we do not need to assume a strong form of introspection to make sense of the use of personal probabilities, the common prior assumption is much stronger. In practice, and more realistically, distinct agents may rather have priors about each other's priors, and could very well be mistaken.

We believe that a distinct model could be designed along those lines, though it would have to be significantly more complex. For our purposes, however, we think it is sufficient to produce a worked out model of the contrast between "around" and "between" along the lines we suggested, setting aside further refinements.

10 Conclusion

In this paper we have pursued two main goals, one broad and one more specific. Our broad goal has been to flesh out the general idea that vague language can be more optimal than precise language in some contexts. One side to that idea is not novel: we find it epitomized in Frazee and Beaver (2010)'s dictum that "vagueness is rational under uncertainty", and in their proposal to substantiate this view in probabilistic and information-theoretical terms. However, another side to it is novel, namely the idea that vagueness can allow a cooperative speaker to achieve an optimal tradeoff between Gricean Quality and Gricean Quantity. To establish this, we have shown that when a speaker is uncertain about the world, the use of a vague preposition like "around" offers in some cases an optimal way to secure Quality (truthfulness) and Quantity (informativeness). That is, we have shown that the use of "around" can be informationally optimal compared to any more precise way for the speaker to convey the information at her disposal (whether by means of exact numerical values or of precise intervals). A critical ingredient of our approach is the idea that vague expressions make it possible for speakers to convey probabilistic information, in a way that precise expressions cannot, without any need to assume that their lexical entry directly refers to probabilities.

Our more specific goal has been to advance our understanding of approximation expressions such as "around". As it turns out, the semantic status of words like "around", "about", "approximately" and "roughly" was put forward by (Wright, 1995, 153-154) as a test case for the epistemic theory of vagueness (Williamson 1994), and so in relation to the broader goal of adjudicating between theories of vagueness. Wright considered as highly objectionable the idea that "roughly 6" should

[25] For a preliminary investigation of rounding in cases in which the speaker is perfectly informed, we refer to Mortier (2019), which develops a model within the Lexical Uncertainty framework of Lassiter and Goodman (2017) where messages that include round numbers are less costly than those using non-round numbers, with the consequence that a fully informed speaker might choose to use an "around"-statement in order to avoid using a non-round number.

have precise but unknowable underlying truth-conditions. Instead, according to Wright "*the role of such particles seems unquestionably to be to introduce some conveniently indeterminate degree of flexibility*" (our emphasis). Wright, however, did not try to produce a conclusive argument in favor of his own view. Instead, he examined objections in support of epistemicism. Wright summarized the epistemic view as follows:

> if I claim that Jones is roughly 6 feet tall, the epistemicist must so construe the truth-conditions of what I say that, for some fixed k and j, they coincide with those of the statement: "Jones is more than $5k$ and less than $6j$ tall".

In support of that view, Wright conceded that when we say "x is roughly 6 feet tall", there must be a maximum distance to 6 such that when x exceed that distance the sentence is no longer true.

Our account of "around" may be seen as vindicating Wright's first intuition. In section 3, we treated "around" as an expression containing an open semantic parameter. From that standpoint, an "around"-statement does not semantically express a proposition unless a value for this parameter is provided, and once a value is provided, the proposition expressed is identical to the one expressed by a certain "between"-statement. We should note, however, that in our model, the listener does not in fact need to pick a value for this parameter in order to interpret an "around"-statement. Rather, what she needs is a probability distribution over this parameter. That is, in our model, the proposition expressed by an "around"-statement relative to a certain fixed value of the parameter does not play any direct role. This means that we could as well directly relativize the meaning of an "around"-statement to a probability distribution over the parameter y, and characterize it directly in terms of an interpretation rule for the literal listener (the one expressed by Equation vii).

In our illustrations, we have assumed that "around n" should not denote an interval larger than $[0, 2n]$, and in section 5.2, we assumed that the context excluded values beyond a certain range.[26] However, it is an essential feature of our model that starting from uniform priors, and given a fixed maximal interval of admissibility for "around", the listener's posterior probability based upon hearing "around" gives non-zero probability to all values within that range. In our framework, then, it does not really make sense to ask what is the 'true' interval denoted by an expression of the form "around n", since the interpretation and use of "around"-statements can be derived without assuming anything substantial about the length of the relevant interval (since we can assume a uniform prior distribution over interval-lengths).

The point we are making here is that our treatment of "around" and similar approximating expressions does indeed guarantee the "conveniently indeterminate degree of flexibility" claimed by Wright, in a way that no truth-conditionally precise surrogate can provide. While we agree with the epistemicist that the use of vague expressions is constrained by general maxims of knowledge

[26] Ferson et al. (2015) provide experimental data showing that when asked to estimate the largest interval compatible with an *around* n-statement, people tend to pick an interval that is much narrower than $[0, 2n]$. Taken at face value, this could suggest that the prior distribution on y categorically excludes too large intervals – since otherwise the posterior distribution resulting from an "around"-sentence would not categorically exclude any value that was not already excluded prior to the utterance. Such a conclusion is however not warranted, and depends on the 'linking' theory that provides the bridge between a specific model and people's behavior in an experimental task. In our setup, even with a uniform prior distribution on the set of intervals of the form $[n - i, n + i]$, with $i \leq n$, as well as on the range of the variable of interest x, the listener's posterior distribution after processing "around n" assigns very low probability to values that are very far from n. It is very plausible that, when asked to estimate an interval, people simply report an interval of values which receive a high enough probability, and therefore exclude values which, without being equal to 0, are in practice negligible.

and rationality, we therefore see the present account as an argument for the irreducibility of the meaning of vague expressions to those of precise expressions.[27] On the epistemicist perspective, vague expressions must be used with a margin of error to forestall error.[28] Here, a vague modifier like "around" is seen rather as a resource to minimize error, but more importantly, as a linguistic means to convey accurate information about one's own state of uncertainty.

Appendix A A limitation result about the LU model

In this appendix, we prove that in the lexical uncertainty model (LU model), if two observations o_1 and o_2 are such that the support of the conditional distributions $P(w|o_1)$ and $P(w|o_2)$ are the same, then, at every level of the recursion, the speaker's probability of using a message m if she observed o_1 is the same as if she observed o_2. It follows that in the LU model, only the *support* of the subjective probability distribution of the speaker, and not its *shape*, plays a role in her choice of a message, in contrast with our model.

The LU model is defined by the following equations, where $[\![m]\!]^i(w)$ is the truth-value of the literal meaning of m, relative to the interpretation function i, in world w, and $[\![m]\!]^i$ denotes the set of worlds where m is true relative to interpretation i (in our setting i is what determines the interpretation of 'around', i.e. a certain value for y). The temperature parameter λ is a non-null, positive real number. For any message m, $c(m)$ is the *cost* of m, a null or positive real number. P is the prior distribution about the possible values of w (world state), o (observation) and i, and is such that the value taken by i is probabilistically independent from w and o.

1. $L^0(w, o|m, i) = \dfrac{P(w, o) \times [\![m]\!]^i(w)}{P([\![m]\!]^i)}$
2. $U^1(m|o, i) = (\sum\limits_w P(w|o) \times \log(L^0(w, o|m, i))) - c(m)$
3. $S^1(m|o, i) = \dfrac{\exp(\lambda U^1(m, o, i))}{\sum\limits_{m'} \exp(\lambda U^1(m', o, i))}$
4. $L^1(w, o|m) = \dfrac{P(w, o) \times \sum\limits_i P(i).S^1(m|o, i))}{\alpha_1(m)}$, where $\alpha_1(m) = \sum\limits_{w',o'} (P(w', o'). \sum\limits_i P(i).S^1(m|o', i)))$
5. For $n \geq 1$, $U^{n+1}(m|o) = (\sum\limits_w P(w|o). \log(L^n(w, o|m)) - c(m)$
6. $S^{n+1}(m|o) = \dfrac{\exp(\lambda U^{n+1}(m, o))}{\sum\limits_{m'} \exp(\lambda U^{n+1}(m', o))}$

[27] Sutton (2018) recently argued that an adequate metasemantics for probabilistic treatment of vagueness is one in which vague expressions do not have truth-conditions proper, but default rules of use. Our account of the meaning of "around" maintains truth-conditions for "around", but as discussed above they do not play any direct role. In our model, the listener, when interpreting an "around"-statement, updates her probability distribution over worlds, but does not exclude any world from the common ground. We leave a more detailed discussion of this aspect, as well as of the rejoinders that could be made on behalf of epistemicism, for another occasion.

[28] See Williamson 1994.

7. For $n \geq 2$, $L^n(w, o|m) = \dfrac{P(w, o) \times S^n(m|o)}{\alpha_n(m)}$, where $\alpha_n(m) = \sum_{w', o'} P(w', o').S^n(m|o')$

Auxiliary Definitions

1. We say that a message m respects Quality with respect to an observation o and an interpretation if, for every w, if $P(w|o) > 0$, then $[\![m]\!]^i(w) = 1$.

2. We say that a message m respects Weak Quality with respect to an observation o if there exists an interpretation i such that $P(i) > 0$ and m respects Quality with respect to o and i.

We will repeatedly use the following facts:

(A-1) **Facts**.

a. If a message m does not respect Quality with respect to an observation o and an interpretation i, then $S^1(m|o, i) = 0$; if m does respect Quality with respect to o and i, then $S^1(m|o, i) > 0$

b. If a message m does not respect Weak Quality with respect to an observation o, then for every $n \geq 2$, $S^n(m|o) = 0$. If m does respect Weak Quality with respect to o, then $S^n(m|o) > 0$.

Proof of the facts in (A-1)

First we prove the result for S^1, then for S^2 and then by induction for higher values of n.

If m does not respect quality with respect to o and i, then for some w such that $P(w|o) > 0$, $[\![m]\!]^i(w) = 0$. For such a w, then, $L^0(w, o|m, i) = 0$, hence $\log(L^0(w, o|m, i)) = -\infty$. So at least one term in the sum which defines $U^1(m|o, i)$ evaluates to $-\infty$, and so the sum itself does, hence $U^1(m|o, i) = -\infty$.[29] Since $S^1(m|o, i)$ involves exponentiating a quantity that is infinitely negative, we have $S^1(m|o, i) = 0$. Reciprocally, if m does respect Quality with respect to o and i, then no term in the sum is infinitely negative, and $U^1(m|o, i)$ is not infinitely negative either, and so $S^1(m|o, i) > 0$

Suppose now that m does not respect Weak Quality with respect to o. Then for every i such that $P(i) > 0$, m does not respect Quality with respect to o, i; and so by the result just proven, every term in the sum $\sum_i P(i).S^1(m|o, i)$ is equal to 0, and so is the sum as a whole. As a result, $L^1(w, o|m) = 0$, for every w. From this it follows that the sum $\sum_w P(w|o). \log(L^1(w, o|m))$ evaluates to $-\infty$, and therefore so does $U^2(m|o)$. $S^2(m|o)$ involves again exponentiating an infinitely negative value, so is equal to 0. Reciprocally, if m respects Weak Quality with respect to o, there is at least one i relative to which $P(i) \times S^1(m|o, i) > 0$, and therefore $\sum_i P(i).S^1(m|o, i)$ is not equal to 0. Since for some w, then, $L^1(w, o|m) > 0$, $U^2(m|o)$ is not infinitely negative, and so $S^2(m|o) > 0$.

Finally, let assume that the result holds for S^n (Induction Hypothesis). Suppose again that m

[29] Strictly speaking, of course, $U^1(m|o, i)$ is simply not defined, since $\log(0)$ is not defined. The point is simply that the limit of $\exp(f(x))$ in 0 is 0 when f diverges to $-\infty$ in 0. Likewise, we also treat the function $x \times \log(x)$ as evaluating to 0 in 0, because even though this function is not defined in 0, its limit in 0 is 0. Here and elsewhere we choose not to introduce explicit reasoning about limits in order to simplify the exposition, with no harmful effects.

does not respect Weak Quality with respect to o. By the Induction Hypothesis, $S^n(m|o) = 0$, and therefore for every w, $L^n(w, o|m) = 0$ (given the definition of L^n). Then $U^{n+1}(m|o) = -\infty$, as in each term of the sum that defines $U^{n+1}(m|o)$, the log-function takes 0 as its argument. It follows that $S^{n+1}(m|o) = 0$. Reciprocally, if m does respect Weak Quality with respect to o, then $S^n(m|o) > 0$, and therefore for some w (w must be such that $P(w, o) > 0$), $L^n(w, o|m) > 0$, from which it follows that $U^{n+1}(m|o)$ is not infinitely negative and therefore that $S^{n+1}(m|o) > 0$.

(A-2) **Lemma**

Let o_1 and o_2 be two observations such that for every w, $P(w|o_1) > 0$ iff $P(w|o_2) > 0$ [we will say that P_{o_1} and P_{o_2} have the same support]. Note that, relative to a given i, m respects Quality with respect to o_1 iff it respects it relative to o_2. If m respects Quality, relative to i, with respect to o_1 and o_2, then the difference $U^1(m|o_2, i) - U^1(m|o_1, i)$ does not depend on m or i, but only on o_1 and o_2 (i.e. it is the same for any m that respects Quality with respect to o_1 and o_2, and i). More specifically:

$$U^1(m|o_2, i) - U^1(m|o_1, i) = \sum_{w \in S} P(w|o_2). \log(P(w, o_2)) - \sum_{w \in S} P(w|o_1). \log(P(w, o_1))$$

(m and i do not appear on the right-hand side).

Proof of the Lemma in (A-2)

Let us assume that m, i and o_1 and o_2 meet the condition stated in (A-2).

We note S the set of worlds such that $P(w|o_1) > 0$ (S is also the set of worlds such that $P(w|o_2) > 0$, by assumption). Note that since m respects Quality with respect to o and i, then if $w \in S$, then $[\![m]\!]^i(w) = 1$

$$\begin{aligned} U^1(m|o_1, i) &= \sum_w P(w|o_1). \log(L^0(w, o_1|m, i)) - c(m) \\ &= \sum_w P(w|o_1). \log\left(\frac{P(w, o_1).[\![m]\!]^i(w)}{P([\![m]\!]^i)}\right) - c(m) \end{aligned}$$

Now, $P(w|o_1) = 0$ if w is not in S and $[\![m]\!]^i(w) = 1$ for any w in S, by Quality. We can therefore continue as follows:

$$U^1(m|o_1, i) = \sum_{w \in S} P(w|o_1). \log\left(\frac{P(w, o_1)}{P([\![m]\!]^i)}\right) - c(m)$$
$$= \sum_{w \in S} P(w|o_1).[\log(P(w, o_1) - \log(P([\![m]\!]^i))] - c(m)$$
$$= \sum_{w \in S} P(w|o_1). \log(P(w, o_1)) - \sum_{w \in S} P(w|o_1). \log(P([\![m]\!]^i)) - c(m)$$
$$= \sum_{w \in S} P(w|o_1). \log(P(w, o_1)) - \log(P([\![m]\!]^i)) - c(m).$$
$$\text{(since } \sum_{w \in S} P(w|o_1). \log([\![m]\!]^i) = \log([\![m]\!]^i) \times \sum_{w \in S} P(w|o_1) = \log([\![m]\!]^i)\text{)}$$

The same formula of course holds for o_2, replacing every occurrence of o_1 with o_2. Given this, when we substract $U^1(m|o_1, i)$ from $U^1(m|o_2, i)$, the terms that depend on m ($-\log(P([\![m]\!]^i)) - c(m)$) disappear, and we get:

$$U^1(m|o_2, i) - U^1(m|o_1, i) = \sum_{w \in S} P(w|o_2). \log(P(w, o_2)) - \sum_{w \in S} P(w|o_1). \log(P(w, o_1))$$

As promised, then, this difference does not depend on m or i.

(A-3) **Lemma**
If two observations o_1 and o_2 are such that the distributions (over w) P_{o_1} and P_{o_2}, defined by $P_o(w) = P(w|o)$, have the same support (i.e for every w, $P_{o_1}(w) > 0$ iff $P_{o_2}(w) > 0$), then, for every interpretation i and every message m, $S^1(m|o_1, i) = S^1(m|o_2, i)$.

Proof of the Lemma in (A-3)

Let o_1 and o_2 be two observations that meet the condition in (A-3).

First consider the case where m does not respect Quality with respect to o_1, o_2, i (again, relative to a fixed i, either it respects quality for both o_1 and o_2, or for neither). In the case, given the first fact in (A-1), $S^1(m|o_1, i) = S^1(m|o_2, i) = 0$.

Consider now the case were m respects Quality with respect to o_1, o_2, i. Again, let S be the set of worlds such that $P(w|o_1) > 0$ (which is also the set of worlds where $P(w|o_2) > 0$) Let us define
$K(o_1, o_2) = \sum_{w \in S} P(w|o_2). \log(P(w, o_2)) - \sum_{w \in S} P(w|o_1). \log(P(w, o_1))$. From the lemma in (A-2), we have: for every m' which respects quality with respect to o_1, o_2, i,

$$U^1(m'|o_2, i) = U^1(m'|o_1, i) + K(o_1, o_2).$$

Now:

$$S^1(m|o_2,i) = \frac{\exp(\lambda.U^1(m|o_2,i))}{\sum\limits_{m'} \exp(\lambda.U^1(m'|o_2,i))}$$

Note that the terms in the sum that constitutes the denominator, of the form $exp(\lambda.U^1(m'|o_2,i))$, are equal to 0 when m' does not respect Quality with respect to o_1, o_2, i. Let us call $\mathcal{M}_{o_1,o_2,i}$ the set of messages that respect Quality with respect to o_1, o_2, i. We then have:

$$\begin{aligned}
S^1(m|o_2,i) == & \frac{\exp(\lambda.U^1(m|o_2,i))}{\sum\limits_{m' \in \mathcal{M}_{o_1,o_2,i}} \exp(\lambda.U^1(m'|o_2,i))} \\
= & \frac{\exp(\lambda.(U^1(m|o_1,i) + K(o_1,o_2)))}{\sum\limits_{m' \in \mathcal{M}_{o_1,o_2,i}} \exp(\lambda.(U^1(m'|o_1,i) + K(o_1,o_2)))} \\
= & \frac{\exp(\lambda.U^1(m|o_1,i)) \times \exp(\lambda.K(o_1,o_2))}{\sum\limits_{m' \in \mathcal{M}_{o_1,o_2,i}} \exp(\lambda.U^1(m'|o_1,i)) \times \exp(\lambda.K(o_1,o_2))} \\
= & \text{ (}\exp(\lambda.K(o_1,o_2))\text{ simplifies)} \\
& \frac{\exp(\lambda.U^1(m|o_1,i))}{\sum\limits_{m' \in \mathcal{M}_{o_1,o_2,i}} \exp(\lambda.U^1(m'|o_1,i))} \\
= & S^1(m|o_1,i)
\end{aligned}$$

(A-4) **Theorem**
Let o_1 and o_2 be such that P_{o_1} and P_{o_2} have the same support, i.e. for every w, $P(w|o_1) > 0 \Leftrightarrow P(w|o_2) > 0$. Then, for any $n \geq 2$, and any message m, $S^n(m|o_2) = S^n(m|o_1)$

Proof of (A-4)

This will be a proof by induction. We first prove the base case, i.e. for $n = 2$.

We start with a counterpart to the Lemma in (A-2):

(A-5) **Lemma**
Let o_1 and o_2 be two observations such that for every w, $P(w|o_1) > 0$ iff $P(w|o_2) > 0$. If m respects Weak Quality relative to both o_1 and o_2, then the difference $U^2(m|o_2) - U^2(m|o_1)$ does not depend on m, but only on o_1 and o_2 (i.e. it is the same for any m that respects Weak Quality with respect to o_1 and o_2).
More specifically:

$$U^2(m|o_2) - U^2(m|o_1) = \sum_{w \in S} P(w|o_2).\log(P(w,o_2)) - \sum_{w \in S} P(w|o_1).\log(P(w,o_1))$$

(m does not appear on the right-hand side).

Proof of the Lemma in (A-5)

Assume that m, o_1 and o_2 meet the condition of the above Lemma. Note that, since P_{o_1} and P_{o_2} have the same support, the interpretations i thanks to which m respects Weak Quality with respect to o_1 are the same as the interpretations i thanks to which m respects Weak Quality with respect to o_2. As before, we call S the support of P_{o_1} and P_{o_2}.
We have, given the definitions:

$$U^2(m|o_1) = \sum_{w \in S} P(w|o_1) \times \log\left(\frac{P(w,o_1).\sum_i P(i).S^1(m|o_1,i)}{\alpha_1(m)}\right)$$
$$= \sum_{w \in S} P(w|o_1) \times [\log(P(w,o_1)) + \log\left(\sum_i P(i).S^1(m|o_1,i)\right) - \log(\alpha_1(m))]$$

Likewise, we have:

$$U^2(m|o_2) = \sum_{w \in S} P(w|o_2) \times [\log(P(w,o_2)) + \log\left(\sum_i P(i).S^1(m|o_2,i)\right) - \log(\alpha_1(m))]$$

Recall that for every i, $S^1(m|o_1,i) = S^1(m|o_2,i)$ (Lemma in (A-3)). It follows that the quantities $\log(\sum_i P(i).S^1(m|o_1,i))$ and $\log(\sum_i P(i).S^1(m|o_2,i))$ are equal. Let us call this quantity K. The fact that m respects Weak Quality with respect to o_1 and o_2 ensures that K is well defined (not infinitely negative, cf. proof of the facts in (A-1)). We can rewrite the above formulae as:

$$U^2(m|o_1) = \sum_{w \in S} P(w|o_1) \times [\log(P(w,o_1)) + K - \log(\alpha_1(m))]$$
$$U^2(m|o_2) = \sum_{w \in S} P(w|o_2) \times [\log(P(w,o_2)) + K - \log(\alpha_1(m))]$$

We have:

$$U^2(m|o_2) - U^2(m|o_1) = (K - \log(\alpha_1(m)) \times \left(\sum_{w \in S} P(w|o_2) - \sum_{w \in S} P(w|o_1)\right)$$
$$+ \sum_{w \in S} P(w|o_2).\log(P(w,o_2)) - \sum_{w \in S} P(w|o_1).\log(P(w,o_1))$$

Now, since $\sum_{w \in S} P(w|o_2) = \sum_{w \in S} P(w|o_1) = 1$, $(\sum_{w \in S} P(w|o_2) - \sum_{w \in S} P(w|o_1)) = 0$, and the above formula simplifies to:

$$U^2(m|o_2) - U^2(m|o_1) = \sum_{w \in S} P(w|o_2).\log(P(w,o_2)) - \sum_{w \in S} P(w|o_1).\log(P(w,o_1))$$

Proof of the base-case ($n = 2$) of the Theorem in (A-4)

Let o_1 and o_2 be such that P_{o_1} and P_{o_2} have the same support. If a certain message m does not satisfy Weak Quality with respect to o_1, o_2, then given the second fact in (A-1), for any $n \geq 2$, $S^n(m|o_1) = S^n(m|o_2) = 0$. We therefore now assume for the rest of the proof that m does satisfy Weak Quality with respect to o_1 and o_2 (recall that it either respects it for both or for neither, because o_1 and o_2 have the same support). Then the proof that $S^2(m|o_1) = S^2(m|o_2)$, with the help of the previous lemma, proceeds in exactly the same way as the proof of the lemma in (A-3) (one just needs to delete i whenever it occurs in the proof of (A-3)).

Inductive step for the Theorem in (A-4)

Let o_1, o_2 be such that they meet the condition of the theorem. We assume that $S^n(m|o_2) = S^n(m|o_1)$. We want to prove that $S^{n+1}(m|o_2) = S^{n+1}(m|o_1)$.

As before, the key intermediate result is the following:

(A-6) **Intermediate result**
Let o_1 and o_2 be two observations such that for every w, $P(w|o_1) > 0$ iff $P(w|o_2) > 0$. If m respects Weak Quality relative to both o_1 and o_2, then the difference $U^{n+1}(m|o_2) - U^{n+1}(m|o_1)$ does not depend on m, but only on o_1 and o_2 (i.e. it is the same for any m that respects Weak Quality with respect to o_1 and o_2). Again, we have:

$$U^{n+1}(m|o_2) - U^{n+1}(m|o_1) = \sum_{w \in S} P(w|o_2). \log(P(w, o_2)) - \sum_{w \in S} P(w|o_1). \log(P(w, o_1))$$

(m does not appear on the right-hand side).

We have:

$$U^{n+1}(m|o_1) = \sum_{w \in S} P(w|o_1) \times \log\left(\frac{P(w, o_1).S^n(m|o_1)}{\alpha_n(m)}\right)$$
$$= \sum_{w \in S} P(w|o_1) \times [\log(P(w, o_1)) + \log(S^n(m|o_1)) - \log(\alpha_n(m))]$$

Likewise, we have:

$$U^{n+1}(m|o_2) = \sum_{w \in S} P(w|o_2) \times [\log(P(w, o_2) + \log(S^n(m|o_2)) - \log(\alpha_n(m))]$$

By the induction hypothesis, $\log(S^n(m|o_2)) = \log(S^n(m|o_1))$. Since m respects Weak Quality with respect to o_1 and o_2, this quantity is well defined (not infinitely negative, cf. proof of the facts in (A-1)). Calling it K, we then have:

$$U^{n+1}(m|o_1) = \sum_{w \in S} P(w|o_1) \times [\log(P(w, o_1)) + K - \log(\alpha_n(m))]$$
$$U^{n+1}(m|o_2) = \sum_{w \in S} P(w|o_2) \times [\log(P(w, o_2)) + K - \log(\alpha_n(m))],$$

The computation then proceeds exactly as in the proof for the Lemma in (A-5).

With the help of this intermediate result, the proof that $S^{n+1}(m|o_2) = S^{n+1}(m|o_1)$ then proceeds exactly in the same way as the proof of the Lemma in (A-3) (again eliminating i everywhere).

Appendix B An alternative model for the literal listener

The model presented in section 3.2 was originally derived from a distinct model of the listener that we first came up with, which we present in this appendix for comparison. Although that model makes basically the same qualitative predictions, it is not a Bayesian model, and it makes different quantitative predictions.

Like the Bayesian model, this model assumes that the listener has a probability distribution P_L over the intervals selected by "around", and over the possible values the variable x of interest might have. However, the alternative model says that the posterior value of x upon hearing "x is around n" is the sum of the conditional probabilities that x takes that value when x belongs to a given interval, weighted by the probability of that interval:[30]

$$P'_L(x = k \mid x \text{ is around } n) =_{df} \sum_i P_L(x = k \mid x \in [n - i, n + i])P_L(y = i) \tag{xiv}$$

To see the difference with the Bayesian model, recall Equation vii, which is reproduced here (with an explicit formula for the proportionality factor):

$$P_L(x = k \mid x \text{ is around } n) = \frac{P_L(x = k) \times \sum\limits_{i \geq |n-k|} P_L(y = i)}{\sum\limits_k P_L(x = k) \times \sum\limits_{i \geq |n-k|} P_L(y = i)} \tag{xv}$$

The two models are distinct. For instance, when P_L is uniform over values of x as well as over candidate meanings for "around", the distribution P'_L obtained is distinct from the one depicted in Figure 2, and it is no longer linear, as represented in Figure 3.

To see more precisely how the two models differ conceptually, the following observation will be useful. Let P_{post} be the posterior probability distribution resulting from updating P_L with the "around n"-message in the Bayesian model, i.e.:

$$P_{post}(x = k, y = i) =_{df} P_L(x = k, y = i \mid x \in [n - y, n + y]).$$

It can be proved that:

$$P_{post}(x = k) = \sum_i P_L(x = k \mid x \in [n - i, n + i])P_{post}(y = i) \tag{xvi}$$

Equation xiv looks almost like xvi, except that the first term in xiv is weighted by the *prior* distribution on the values of y (the candidate meanings for "around") instead of the posterior. This is the sense in which the model proposed in xiv is not Bayesian. Instead of the listener updating also her probability of intervals after hearing "x is around n", the listener does not make her interval

[30] This characterization of the listener happens to be identical (modulo differences in notations) to one that is discussed in Appendix B of Bergen et al. (2016), statement (41).

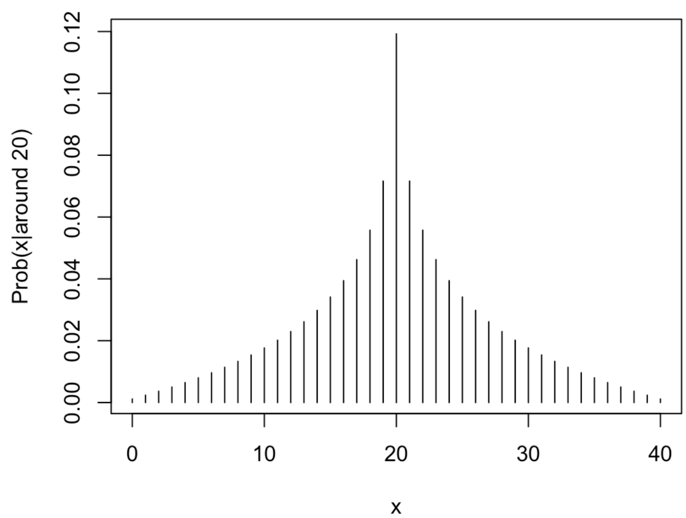

Figure 3: Probability P'_L of $x = k$ when hearing "x is around 20", with maximum interval $[0, 40]$

probability depend on that information. The model is not illegitimate for that matter. Regarding our explananda, it makes the same central prediction: this model can be used to derive the Ratio Inequality. If we used this model in order to characterize the literal Listener, we could still build an RSA model, in the same way as we did in section 6, and we would derive qualitatively similar predictions.

The proof of xvi goes as follows. P_{post} being a probability distribution, it satisfies, for any k:

$$P_{post}(x = k) = \sum_i [P_{post}(x = k \mid y = i) \times P_{post}(y = i)] \tag{xvii}$$

Let us develop the first factor in the sum, $P_{post}(x = k \mid y = i)$.
Since, in general, $P_C(A|B) = P(A|B \wedge C)$ [where P_C is P conditionalized on event C] we have:

$$\begin{aligned} P_{post}(x = k \mid y = i) &= P_L(x = k \mid y = i \wedge x \in [n - y, n + y]) && \text{(xviii)} \\ &= \frac{P_L(x = k \wedge y = i \wedge x \in [n - y, n + y])}{P_L(y = i \wedge x \in [n - y, n + y])} && \text{(xix)} \\ &= \frac{P_L(x = k \wedge y = i \wedge x \in [n - i, n + i])}{P_L(y = i \wedge x \in [n - i, n + i])} && \text{(xx)} \end{aligned}$$

Since the random variables x and y are independent, the events $\ulcorner y = i \urcorner$ and $\ulcorner x = k \wedge x \in [n - i, n + i] \urcorner$ are independent, thanks to which we can simplify the formula above:

$$\ldots = \frac{P_L(y=i) \times P_L(x=k \wedge x \in [n-i, n+i])}{P_L(y=i) \times P_L(x \in [n-i, n+i])} \tag{xxi}$$

$$= \frac{P_L(x=k \wedge x \in [n-i, n+i])}{P_L(x \in [n-i, n+i])} \tag{xxii}$$

$$= P_L(x=k \mid x \in [n-i, n+i]) \tag{xxiii}$$

Plugging this last formula into xvii, we get:

$$P_{post}(x=k) = \sum_i [P_L(x=k \mid x \in [n-i, n+i]) \times P_{post}(y=i)] \tag{xxiv}$$

Appendix C Two alternative RSA models (discussed in section 8)

C.1 A variant of our model which uses the standard utility function

In our official model, the utility function for the speaker is defined by the following equations, where P_o is understood to be the posterior distribution over the variable of interest (here notated w, for *world*) induced by observation o, and L^n_m is the posterior distribution over w of the level-n listener who has processed a message m. These distributions, importantly, are not joint distributions over w and o.

$$U^{n+1}(m, o_j) = -D_{KL}(P_o || L^n_m)$$

Developing the formula for KL-divergence, this is more explicitly cashed out as:

$$\begin{aligned} U^n(m, o) &= \sum_w P(w|o) \times [\log(L^n(w|m)) - \log(P(w|o))] \\ &= \sum_w P(w|o) \times [\log(\sum_{o'} L^n(w, o'|m)) - log(P(w|o))] \\ &= \sum_w P(w|o) \times \log(\sum_{o'} L^n(w, o'|m)) - \sum_w P(w|o) \times \log(P(w|o)) \end{aligned}$$

Note that that the second term, $-\sum_w P(w|o) \times \log(P(w|o))$, does not depend on the message m. For this reason it can be dropped: dropping this term amounts to adding a constant term to the utility of each message (relative to a fixed observation o), which has no effect when we apply the *SoftMax* function in order to derive the speaker's behavior. So we can as well use the following utility function, with no change whatsoever in the behavior of the model:

$$U^n(m, o) = \sum_w P(w|o) \times \log(\sum_{o'} L^n(w, o'|m))$$

Now, we can also consider a model whose general architecture is like ours, where the literal listener L^0, in particular, is exactly the same as the one we defined, but where we use the standard utility function of the RSA framework, which is based on the KL-divergence between the joint distribution over (w, o) of the level-n listener, and the joint distribution of the speaker which results from an observation o (such a joint distribution assigns probability 0 to all pairs (w, o') where $o' \neq o$, i.e. $P(w, o'|o) = P(w|o)$ if $o' = o$, otherwise $P(w, o'|o) = 0$).

This amounts to moving to the following utility function, which is the standard one in the RSA framework (ignoring the cost term):

$$U^n(m, o) = \sum_w P(w|o) \times \log(L^n(w, o|m))$$

Keeping all the other ingredients of the model presented in sections 6 and 7, we obtain, with such a model, numerically different results from those of our main model, but qualitatively similar ones, in the following sense: the pragmatic speaker (at different recursive depths) can have a preference for an "around"-statement over any "between"-statement in some situations where there are able to exclude the peripheral values 0 and 8 (and so could have said, e.g., *between 1 and 7*) but have a private distribution that is strongly biased towards values closer to 4. Crucially, for this model, the limitation result proved in Appendix A for the standard LU model does not hold.

C.2 A variant of the LU model where the utility function is as in our own model

We notate $[\![m]\!]^i(w)$ the truth-value of the literal meaning of a message m, relative to the interpretation function i, in world w, and $[\![m]\!]^i$ denotes the set of worlds where m is true relative to interpretation i (in our setting i is what determined the interpretation of 'around', i.e. a certain value for y). The temperature parameter λ is a non-null, positive real number. For any message m, $c(m)$ is the *cost* of m, a null or positive real number. P is the prior joint distribution on worlds and observations.

Below we present the equations that characterize the modified LU model. The crucial difference with the standard LU model shows up in the utility functions (lines 2 and 5), where $L^0(w, o|m, i)$ and $L^n(w, o|m)$ have been re replaced, respectively, with $L^0(w|m, i)$ and $L^n(w|m, i)$, which are themselves equal, respectively, to $\sum_{o'} L^0(w, o'|m, i)$ and $\sum_{o'} L^n(w, o'|m)$.

Importantly, the limitation result reported in Appendix A for the standard LU model no longer holds for this model.

1. $L^0(w, o|m, i) = \dfrac{P(w, o) \times [\![m]\!]^i(w)}{P([\![m]\!]^i)}$
2. $U^1(m|o, i) = (\sum_w P(w|o) \times \log(L^0(w|m, i))) - c(m)$
   $= (\sum_w [P(w|o) \times \log(\sum_{o'} L^0(w, o'|m, i))]) - c(m)$
3. $S^1(m|o, i) = \dfrac{\exp(\lambda U^1(m, o, i))}{\sum_{m'} \exp(\lambda U^1(m', o, i))}$
4. $L^1(w, o|m) = \dfrac{P(w, o) \times \sum_i P(i).S^1(m|o, i))}{\alpha_1(m)}$, where $\alpha_1(m) = \sum_{w', o'} (P(w', o'). \sum_i P(i).S^1(m|o', i)))$

5. For $n \geq 1$, $U^{n+1}(m|o) = (\sum_{w} P(w|o). \log(L^n(w|m)) - c(m)$
   $= (\sum_{w}[P(w|o) \times \log(\sum_{o'} L^n(w, o'|m))] - c(m)$

6. $S^{n+1}(m|o) = \dfrac{\exp(\lambda U^{n+1}(m, o))}{\sum_{m'} \exp(\lambda U^{n+1}(m', o))}$

7. For $n \geq 2$, $L^n(w, o|m) = \dfrac{P(w, o) \times S^n(m|o)}{\alpha_n(m)}$, where $\alpha_n(m) = \sum_{w', o'} P(w', o').S^n(m|o')$

References

Bergen, L., Levy, R., and Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Borel, E. (1907). Un paradoxe économique: le sophisme du tas de blé et les vérités statistiques. *La Revue du Mois*, 4:688–699. English translation by P. Égré and E. Gray, in *Erkenntnis* (79), 1081-1088, 2014: An economic paradox: the sophism of the heap of wheat and statistical truths.

Channell, J. (1985). Vagueness as a conversational strategy. *Nottingham Linguistic Circular*, pages 3–24.

Cobreros, P., Egré, P., Ripley, D., and van Rooij, R. (2012). Tolerant, classical, strict. *The Journal of Philosophical Logic*, 41(2):347–385.

van Deemter, K. (2009). Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, 38(6):607.

Égré, P. and Barberousse, A. (2014). Borel on the Heap. *Erkenntnis*, 79:1043–1079.

Égré, P. and Icard, B. (2018). Lying and vagueness. In Meibauer, J., editor, *The Oxford Handbook on Lying*, pages 354–369. Oxford University Press.

Égré, P., Ripley, D., and Verheyen, S. (2019). The sorites paradox in psychology. In Oms, S. and Zardini, E., editors, *The Sorites Paradox*, pages 263–286. Cambridge University Press.

Ferson, S., O'Rawe, J., Antonenko, A., Siegrist, J., Mickley, J., Luhmann, C. C., Sentz, K., and Finkel, A. M. (2015). Natural language of uncertainty: numeric hedge words. *International Journal of Approximate Reasoning*, 57:19–39.

Frank, M., Goodman, N., Lai, P., and Tenenbaum, J. (2009). Informative communication in word production and word learning. In *Proceedings of the annual meeting of the Cognitive Science Society*, number 31.

Frazee, J. and Beaver, D. (2010). Vagueness is rational under uncertainty. In Aloni, M., Bastiaanse, H., de Jager, T., and Schulz, K., editors, *Logic, Language and Meaning. Lecture Notes in Computer Science*. Springer.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Grice, H. P. (1989 [1967]). Logic and conversation. In *Studies in the Way of Words*. Harvard University Press.

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

Krifka, M. (2007). Approximate interpretation of number words. In Bouma, G., Krämer, I., and Zwarts, J., editors, *Cognitive Foundations of Communication*, pages 111–126.

Lassiter, D. and Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC Press. 1st Edition.

Mortier, A. (2019). *Semantics and Pragmatics of Approximation Expressions.* Master's thesis, ENS, under the supervision of Paul Égré and Benjamin Spector.

Russell, B. (1923). Vagueness. *The Australasian Journal of Psychology and Philosophy*, 1(2):84–92.

Smith, N. J. J. (2008). *Vagueness and Degrees of Truth.* Oxford University Press, Oxford.

Solt, S. (2014). An alternative theory of imprecision. In *Semantics and Linguistic Theory*, volume 24, pages 514–533.

Sutton, P. R. (2018). Probabilistic approaches to vagueness and semantic competency. *Erkenntnis*, 83(4):711–740.

Veltman, F. (2001). Het verschil tussen *vaag* en *niet precies.* University of Amsterdam.

Verheyen, S., Dewil, S., and Égré, P. (2018). Subjectivity in gradable adjectives: The case of *tall* and *heavy*. *Mind & Language*, 33(5):460–479.

Williamson, T. (1994). *Vagueness.* Routledge, London.

Wright, C. (1995). The epistemic conception of vagueness. *The Southern Journal of Philosophy*, 33(S1):133–160.

Declaration of contribution: PE and BS share first authorship in this paper. PE came up with the core idea of the paper, which he then developed with SV, AM, and BS. PE and SV originally proposed the model in Appendix B, and PE and BS later worked out its Bayesian counterpart. BS, AM and PE identified the ratio inequality, and AM investigated the model by PE-SV. BS is responsible for the speaker model and the RSA implementation, and for the comparison with the Lexical Uncertainy model. PE and BS jointly wrote the first complete draft of the paper, which the four authors finalized together.