Precisifying approximation: *Around vs Between*

Adèle Hénot-Mortier (MIT), joint work with Paul Égré (ENS), Benjamin Spector (ENS) & Steven Verheyen (EUR)

Research question

Numerical approximation expressions, such as *around n* and *between x and y*, convey some degree of uncertainty about an exact numerical value. As Ex. (1) shows, *around n* and *between x and y* sound roughly equivalent for well-adjusted values of n, x and y.

(1) a. *Around* 20 people came to the party.
b. *Between* 15 and 25 people came to the party.

Those expressions however, differ in two points: **(1)** *Around* emphasizes a "central" value while *between* emphasizes upper and lower bounds;[1] **(2)** *around* seems more felicitous than *between* in a context where the exact number is known but there is no pressure to be precise. This suggests that *around* is vague in a sense that *between* is not. We devise a Bayesian model of those two approximators, which predicts that *around* conveys some extra information about the distribution of possible numbers, which causes the posterior (computed upon hearing *around n*) to be more "peaked" than the prior. We verify this prediction by conducting an online experiment on Amazon Mechanical Turk.

Basic semantics of *around* and *between*

We define a basic, non-probabilistic semantics for *around* and *between*, understood as generalized quantifiers.

$$[\![\textit{between x and y}]\!](P)(Q) = |\iota_{\text{Max}}\, X.\, P(X) \wedge Q(X)| \in [x; y] \quad (1)$$

Eq. (1) says that the maximum plurality verifying both the restrictor and nuclear scope (let us call this plurality the *witness*), has its cardinal included in $[x; y]$. Ex. (1b) then means that the maximum plurality of people who came to the party has k members, with $k \in [x; y]$. We now define an entry for *around n*, inspired from that of gradable adjectives (Kennedy, 2007):

$$\begin{aligned}[\![\textit{around n}]\!](P)(Q) &= |\iota_{\text{Max}}\, X.\, P(X) \wedge Q(X)| \in [n \pm i] \\ &= [\![\textit{between n-i and n+i}]\!](P)(Q) \quad (2)\end{aligned}$$

Eq. (2) says that the size of the witness is included in an interval centered on n and of total size $2i$+1. We take i to be a free variable, which depends on the order of magnitude of n,[2] its granularity, and the preciseness of the QUD.[3] The crucial difference between *around* and *between* is taken to be this free variable i, that the listener has to guess, along with the exact number. This makes *around* inherently vague and motivates a Bayesian analysis of numerical approximators, in the spirit of (Lassiter and Goodman, 2013; Qing and Franke, 2015; Bergen et al., 2016).

A simple Bayesian Model of numerical approximation and its prediction

We consider a conversation focused on a specific QUD, e.g., *How many people came to the party?* The speaker $\mathcal{S}$ (resp. the listener $\mathcal{L}$) has a prior distribution $\mathcal{P}_\mathcal{S}$ (resp. $\mathcal{P}_\mathcal{L}$) on exact numerical values that constitute best answers to the QUD. $\mathcal{S}$ tries to communicate $\mathcal{P}_\mathcal{S}$ to $\mathcal{L}$, by uttering either *around n* or *between x and y* (Egré et al., 2020). Optimal values for n or x and y are obtained by making the assumption that $\mathcal{L}$ updates $\mathcal{P}_\mathcal{L}$ in a Bayesian way.

Between-update. We assume that it is uniform and rules out the values that are not in $[x; y]$. All numbers within $[x; y]$ become equally more probable, i.e. sheer uncertainty is being conveyed:

$$\mathcal{P}_\mathcal{L}[k \mid \textit{between x and y}] \propto \begin{cases} \mathcal{P}_\mathcal{L}[k] & \text{if } k \in [x; y] \\ 0 & \text{if } k \notin [x; y] \end{cases} \quad (3)$$

Around-update. Recall Eq. (2) which states that *around n* is equivalent to *between n-i and n+i* for a given i. Since $\mathcal{L}$ is not certain of the i that $\mathcal{S}$ is using ($\mathcal{S}$ might not be either!), the *around*-update amounts to marginalizing on all possible $[n \pm i]$ intervals (for simplicity we assume here that those are equiprobable).

$$\begin{aligned}\mathcal{P}_\mathcal{L}[k \mid \textit{around n}] &\propto \sum_i \mathcal{P}_\mathcal{L}[k \mid \textit{between n-i and n+i}] \\ &\propto \sum_{i=|n-k|}^{n} \frac{\mathcal{P}_\mathcal{L}[k]}{\sum_{j=n-i}^{n+i} \mathcal{P}_\mathcal{L}[j]} \quad (4)\end{aligned}$$

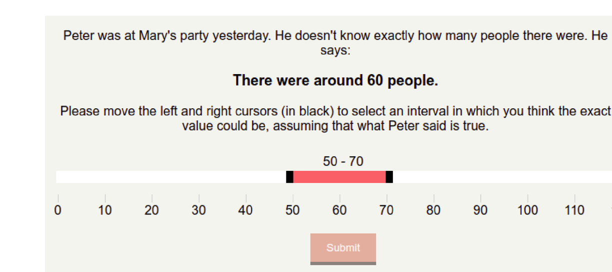
This update does *not* redistribute the probabilistic mass from the prior distribution *equally* among the possible numbers. This suggests that *around*, unlike *between*, conveys some extra information about what $\mathcal{S}$ thinks the most probable values are (Egré et al., 2020). More specifically, it can be shown that the *around*-update systematically gives more mass to values that are closer to n. Let k, k' be s.t. $|n-k| < |n-k'|$:

$$\frac{\mathcal{P}_\mathcal{L}[k \mid \textit{around n}]}{\mathcal{P}_\mathcal{L}[k' \mid \textit{around n}]} \geq \frac{\mathcal{P}_\mathcal{L}[k]}{\mathcal{P}_\mathcal{L}[k']} \stackrel{4}{=} \frac{\mathcal{P}_\mathcal{L}[k \mid \textit{between x and y}]}{\mathcal{P}_\mathcal{L}[k' \mid \textit{between x and y}]} \quad (5)$$

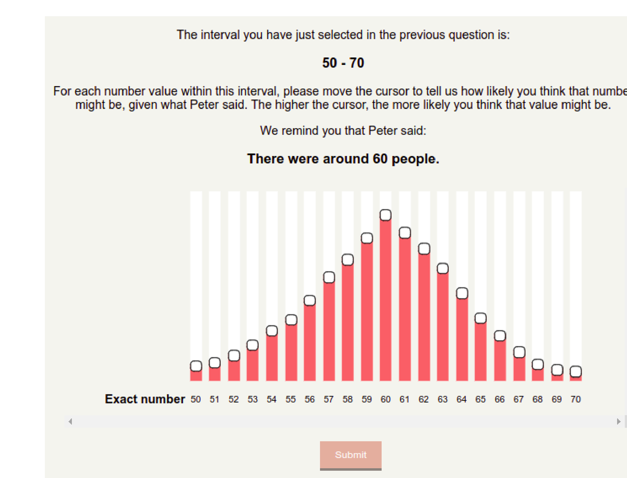
Experiment

Design. We conducted an online experiment inspired from (Channell, 1994) on 145 participants to test Eq. (5). The goal was to elicit posteriors for *around n* and *between x and y*, for each participant (within-subject design). Given a certain value of n, x and y were determined as a function of the participant's answer to the *around n* stimulus, s.t. the elicited *around*- and *between*-distributions form near-minimal pairs. For each participant, two ratios similar to those in Eq. (5) were computed from the elicited distributions.[5] The hypothesis was that, for a given participant, the *around*-ratio should be higher than the *between*-ratio.

Task. During the **Interval** task (Fig. (1a)), the participant was presented with a sentence of the form *Around n / Between x and y people came to the party* (n=40, 50 or 60, randomized across participants) and was asked to give the interval within which the exact value should be. The first Interval task was testing *around n*, and the bounds returned by the participant for this task would be subsequently used as the x and y values in *between x and y*. This ensured that each participant was tested on approximation expressions that were, in their opinion, optimally close in meaning. During the **Histogram** task (Fig. (1b)), the participant had to assign weights to each number within the interval they had given in the Interval task. The weights were supposed to reflect the "likelihood" of each number.



(a) Interval task



(b) Histogram task

Figure 1. Screen captures of the two tasks (n randomly set to 60)

Screen	Expression	Task	Expected output
1	*Around 50*	Interval	$[a; b]$ with $a \leq 50 \leq b$
2	*Around 50*	Histogram	$\mathcal{P}_\mathcal{L}[k \mid \textit{around 50}]$
3	*Between a and b*	Interval	$[c; d]$ with $c = a$, $d = b$
4	*Between a and b*	Histogram	$\mathcal{P}_\mathcal{L}[k \mid \textit{between a and b}]$

Table 1. Arrangement of the tasks (n randomly set to 50)

Results & Discussion

Results. We performed a Sign test on the pairs of ratios obtained for each participant, which proved significant ($p = 4.99 \times 10^{-8}$). The effect was of medium size (Cohen's $d = 0.497$), and Bonferroni-corrected Wilcoxon tests performed on each number group confirmed its significance. This suggests that the hypothesis according to which an *around*-update moves the probability mass towards central values more than a *between*-update does, cannot be rejected.

Discussion. The result is encouraging but the design suffers from one major flaw, namely, the absence of randomization in the order of presentation of *around* vs *between*. Since the x and y values of the *between x and y* expression presented to the participant were determined during the Interval task of *around*, this task *had* to appear *before* the Interval task of *between*. This limit of the design may be the cause of non-counterbalanced order effects. A solution may be to reorder the trials from Tab. (1) in the following way: start with screens 1 and 3 (Interval tasks) in that order, to determine x and y. Then, present screen 2 and 4 (the most critical Histogram tasks) in a *randomized order*.[6]

Conclusion

We devised a Bayesian model of *around* and *between*, which makes a clear prediction about the effect of those expressions on the distribution of possible number values. *Between x and y* conveys sheer uncertainty about the numbers within its bounds, which leads to a "uniform" update. Surprisingly perhaps, *around n* ends up communicating useful information about which values should be seen as the most probable (namely, those closer to n). This prediction is in line with the claim that using vagueness can be rational under uncertainty (Frazee and Beaver, 2010). We conducted an experiment which confirmed this claim, despite one *caveat* (order effects), that we are currently trying to fix.

Selected references

Channell, J. (1994). *Vague language*. Oxford University Press. ★ Dehaene, S. (2003). The neural basis of the weber–fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*. ★ Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *L&P*. ★ Frazee, J., & Beaver, D. (2010). Vagueness is rational under uncertainty. In *Logic, language and meaning*. Springer. ★ Cummins, C., Sauerland, U., & Solt, S. (2012). Granularity and scalar implicature in numerical expressions. *L&P*. ★ Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*. ★ Qing, C., & Franke, M. (2015). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. *Semantics and Linguistic Theory*. ★ Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*. ★ Égré, P. (2016). Vague judgment: A probabilistic account. *Synthese*. ★ Egré, P., Spector, B., Mortier, A., & Verheyen, S. (2020). On the optimality of vagueness: "around", "between", and the gricean maxims. arXiv. ★

[1] Relatedly, *around* is sorites-susceptible while *between* is not.

[2] i usually increases with the value of n. See (Égré, 2016) for a related example; the phenomenon is perhaps linked to a robust observation made in the field of numerical cognition (Dehaene, 2003 i.a.), in support of a logarithmic representation of the numerical scale.

[3] i usually increases with the coarseness of n, modulated by the coarseness of the QUD. See (Cummins et al., 2012) for a related example and experimental evidence supporting this claim.

[4] The equality holds for x and y s.t. $x \leq k, k' \leq y$, i.e. s.t. Case 1. of Eq. (3) applies to k and k'.

[5] In practice, a "grand" ratio was computed for each approximator by averaging the ratios obtained for all possible pairs (k, k') (1) s.t. $|n-k| < |n-k'|$; (2) present in the support of both the *around*- and the *between*-distribution, and (3) not made salient by the experimental design (i.e., different from n, x or y).

[6] We thank Athulya Aravind for this very clever suggestion, that we are currently implementing in a follow-up experiment.

mortier@mit.edu