Talking about probability: approximately optimal thresholds for *probable*

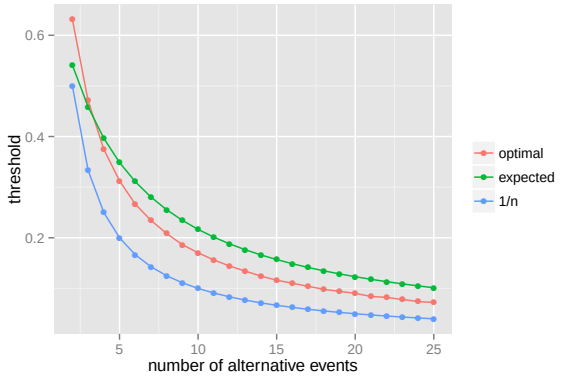
Probability expressions *probable* and *likely* communicate a speaker's level of credence in a proposition of interest. Recently, a degree-based semantics has been proposed, inspired by semantics for relative gradable adjectives (Yalcin, 2010; Lassiter, 2011): an event is *probable/likely* if its probability exceeds a certain contextually variable threshold. One reason for such a move, is the amply demonstrated "alternative outcome effects" (e.g. Windschitl and Wells, 1998): acceptability judgements of "probable/likely E" vary with the size n of the set of contextually salient alternative events of which E is a member. Here, we take inspiration from recent like-minded approaches to determining thresholds for gradable adjectives (Lassiter and Goodman, 2014; Qing and Franke, 2014) and address the question what a pragmatically optimal threshold for *probable/likely* would be from a pragmatic point of view, as a function of n. We ask and answer provisionally what an optimal lexical meaning for *probable/likely* would be if its meaning evolved in order to be communicatively efficient. (There are subtle differences between *probable* and *likely*, but these must be ignored at the high level of abstraction that we consider here; we therefore mention only *probable* henceforth.) The predictions of the proposed model are corroborated by data from a truth-value judgement task. The data reveal that the abstract model is a better predictor than a naive straw-man semantics that assumes that "probable E" is true iff E is more likely than $1/n$.

Model. To assess whether something is optimal, we need to supply an implicit argument: "optimal *for what*?" We assume here that the predominant function of statement "probable E" is to communicate the speaker's degree of belief $P_S(E) = p^*$, where E is one out of n alternative outcomes. More concretely, we assume that the speaker wants to induce in the listener a belief about event E that is at least as high as his own, or otherwise one that comes close: if $p_L = P_L(E)$ is the listener's belief, the utility of the speaker is $\mathrm{U}(p^*, p_L) = 1$ if $p* \leq p_L$ and $1 - (p^*, p_L)^2$ otherwise.

Following mentioned previous work on gradable adjectives, we assume that the speaker faces a generic interpreter to whom he can either say "probable E" or nothing at all. From an evolutionary point of view, we want to know the expectation of the interpretation that listeners would form, if a threshold θ is imagined as established. If any listener belief is a priori equally likely to be encountered, these are simply the expectations under unbiased Dirichlet distributions, either unconditioned (in case of silence) or conditioned on $P_L(E) \geq \theta$. This gives us expectations about listener reactions: $p_L = \frac{1}{n}$ if the speaker says nothing, and $p_L = \frac{1-\theta}{n} + \theta$ if the speaker utters "probable E." We can then determine the utility of having a fixed semantic convention θ for *probable* as a function of n, for a speaker with hypothetically fixed p^*: $\mathrm{U}(p^*, \theta, n) = \mathrm{U}(p^*, \frac{1}{n})$ if $p^* < \theta$; and $\mathrm{U}(p^*, \frac{1-\theta}{n} + \theta)$ otherwise. The expected utility of having a conventional threshold θ is then obtained by weighing in the probability $P(p^*)$ of how likely a speaker, on average, would have a belief value p^* for an arbitrary event E when there are n possible outcomes: $\mathrm{EU}(\theta, n) = \int_0^1 P(p^*) \cdot \mathrm{U}(p^*, \theta, n)\ \mathrm{d}p^*$. If any speaker belief is equally likely to occur, the probability $P(p^*)$ can be analytically determined. It is proportional to the volume of the $(n-2)$-dimensional regular simplex with side length $1 - p^*$ (details omitted for abstract).

This gives crisp predictions about optimal thresholds for *probable* as a function of n, under the assumed pragmatic function of communicating degrees of belief. The figure below shows the optimal thresholds for different n (red). To make reasonable empirical predictions, we would not want to assume that speakers strictly conform to the optimal threshold, but only that they do so in tendency. Expectations for a hypothetical population of speakers that would use a threshold with a probability proportional to its expected utility are also plotted (green), as are the predictions of the naive benchmark approach that sets the threshold to $\frac{1}{n}$ (blue).

Empirical data. We would like to compare empirically the predicted expectations of approximately optimal thresholds (green) to the "naive approach" (blue). We focus on moderately small n, because we hypothesize that subjects might be prone to resort to more parsimonious conceptualizations of the event space for larger n, an issue that has to be left open for further testing.



We collected truth-value judgements of 425 participants on MTurk, giving us 25 judgements per relevant condition. Participants were first presented with a background scenario of a fictitious game show in which $n \in \{2, 3, 4, 5\}$ candidates answer questions. For each correctly answered question, the game master adds balls with the name of the candidate who answered correctly to an urn. At the end of the question phase, the game master draws a ball from the urn to determine the winner. Participants were asked about the truth value of sentences "The winner is possibly/probably/certainly candidate A" for varying numbers of balls for candidate A. There were always 100 balls in total and all candidates other than A had equal (or almost equal, due to rounding) numbers of balls. We removed data points of participants who failed to give the logically expected (slack-free) response for both *possibly* and *certainly*. Proportions of remaining truth-value judgments for *probable* are shown in the figure at the bottom. As expected, our data reflect the observed "alternative outcome effect": e.g., with fixed probability for A's winning at .4, the acceptability of the target statement increases as n increases (detailed descriptive statistics omitted for abstract).

The figure also shows the predictions of two parameterized models. One takes the expected approximately optimal thresholds from the figure above as input, the other relies on the naive $\frac{1}{n}$ thresholds. In either case, we assume that subjects apply thresholds noisily. Concretely, if a model's predicted threshold is θ, the probability that subjects use a threshold value x is assumed to be given by a normal distribution with mean θ and standard deviation σ as a free parameter that captures the uncertainty about / vagueness of θ in either model. We determined the values of σ that best approximated the models' predictions to the data (by minimizing residuals). The models' best predictions are plotted in the figure below. It is even visually apparent that the "naive" model makes worse predictions. The correlation of prediction and observation points is 0.88 for the naive model, and 0.94 under the theory-driven model, suggesting that the latter, despite its high-level of abstraction, makes astonishingly accurate empirical predictions.

References. Lassiter (2011), PhD Thesis. Lassiter & Goodman (2014), SALT 23. Qing & Franke (2014), SALT 24. Windschitl & Wells, J. Personality & Social Psychology. Yalcin (2010), Philosophy Compass.