# Hierarchical Bayesian Sampling of Frequency Adverbs

**Jeremy Ma**
Massachusetts Institute of Technology
jma22@mit.edu

**Alex Abate**
Massachusetts Institute of Technology
aabate@mit.edu

## Abstract

Uncertainty is built into the nature of our world, but what capacity does language have to transmit information about probabilities? Is there a mathematical framework which can model the human-interpreted uncertainties encoded in frequency adverbs? To answer these questions, we first developed a framework for understanding the way humans interpret frequency adverbs. Next, we applied a hierarchical Bayesian structure to model the effects of uncertainty modifiers in high-order probabilistic phrases (those with multiple frequency adverbs). Finally, we experimentally tested our model against psychophysical data. Our results demonstrate the validity of our mathematical notion of nested frequency adverb structures and the decomposability of high-order phrases. Furthermore, they shed light into the internal sampling system humans use in natural language processing.

**Keywords:** frequency adverbs, Bayesian Hierarchical Modeling

## Introduction

Many events in life are probabilistic; for example, whether it would rain tomorrow or whether the train would get to the station on time. One challenging task of human language is to convey the probabilities of these events in words, despite the fact that the exact probabilities of such events are usually unavailable to the speaker. As a result, probabilistic descriptors in natural language tend to have a measure of uncertainty along with it. For example, given the statement ”the train often arrives on time”, one might guess that 81 out of 100 trains arrive on time, however the guess that 90 out of 100 trains arrive on time is also not invalid. It is one task for the speaker to select a descriptor that matches the probability that they want to convey. The more difficult task is on the listener since one would have to arrive at an estimate of a probability given only the mere descriptors of such a probability.

Even more elusive are the meanings of high-order probabilistic statements, or those with multiple adverbs (for example, ”the train is occasionally never late”), where language's nested structure allows the first adverb to modify the uncertainty associated with the second adverb. Is there a precise mathematical formulation of such interactions? How do humans interpret uncertainty in the context of multiple adverbs? Can a conditional Bayesian framework account for the internal conception this uncertainty?

To study these questions, we sought to develop a model for the distributions governing words and then compare our predictions to human psychophysical data.

## Relation to previous work

Previous work has shown that we can probabilistic frame generics with a Bayesian framework (Tessler & Goodman, 2018). Further, (Herbstritt & Franke, 2019) examined a Bayesian computational model of probabilistic expressions, demonstrating that the meaning of probabilistic statements, such as probably and likely, (and second-order descriptions of those statements) could be adequately captured by a mathematical framework. This study extends the idea of modelling our expressions to words that express the frequency of events as a Bayesian computational model.

There are couple distinctions between this paper and previous studies. Firstly, this paper aims to look at frequency adverbs instead of probabilistic statements. This means that the event being described has to be a Bernoulli event parameterizable by a single $\theta$

Secondly, most papers uses threshold-based semantics to determine the truth of statements(Schöller & Franke, 2017), we focus on a sampling-based approach. This means that the parameters of interest are not simply thresholded in value, but rather sampled from some distribution. This means that the statement ”event A always happens” does not simply mean ”event A happens more than 99% of the time”, but rather ”event A happens around n% of the time”, with n being some possibly normal distribution with mean 99%. By using a sampling-based approach, this allows more freedom and flexibility in the expression. However, these expressions could have a higher order as stacking descriptors such as ”never always” is perfectly valid and has it's semantic purpose (Moss, 2015). By having a richer representation of first order expressions, this allows a more detailed formulation of second-order descriptions.

Some key assumptions we make in this study are that humans draw samples from a distribution when interpreting likelihood (rather than simply finding the maximum-likelihood event), everyone of a particular cultural background samples from the same distribution, and that distributions defined by high-order probabilistic statements can be decomposed into their respective parts.

## Formulation

First let us look at descriptions regarding events. The probability statement ”Event A rarely never happens” could be broken down into two main components - the event of interest ”A”, and the probability descriptor ”rarely never”. In this study we would focus on frequency adverbs, specifically the following six - ”never”, ”rarely”, ”occasionally”, ”often”, ”usually”, ”always”.

Frequency adverbs are a subset of probability descriptors which include non-frequency adverbs such as ”probably” and ”certainly”. One special property of frequency adverbs is that it is only used to describe Bernoulli events, as frequencies could only be attributed to events that either happens, or not.

Therefore, the frequency of event A happening out of some number of trials could be described using frequency adverbs.

## Event Frequency

Let us formally define event A. Event A is distributed according to a Bernoulli distribution parameterized by $\theta_A$, which is the probability of event A happening. Specifically:

$$p_{A|\theta_A}(A|\theta_A) = \begin{cases} \theta_A & \text{if } A = 1 \\ 1-\theta_A & \text{if } A = 0 \end{cases} \tag{1}$$

However, the exact value of $\theta_A$ is usually unavailable, so there would be some uncertainty regarding $\theta_A$ and a therefore a distribution of $p_{\theta_A}(\theta_A)\ \forall 0 \leq \theta_A \leq 1$. The probability of event A happening with an uncertain $\theta_A$ is:

$$\begin{aligned} p_A(A=1) &= \int_0^1 p_{A|\theta_A}(A=1|\theta_A) p_{\theta_A}(\theta_A) d\theta_A \\ &= \int_0^1 \theta_A p_{\theta_A}(\theta_A) d\theta_A \\ &= E_{p_{\theta_A}}[\theta_A] \end{aligned} \tag{2}$$

This shows that the probability of event A happening would instead be the expected value of $\theta_A$. From the point of view of a listener receiving a probability statement on event A, as the only source of information is the descriptors, the source of uncertainty also comes from the use of descriptors and therefore each descriptor uniquely specifies, or parameterizes, the distribution $p_{\theta_A}(\theta_A)$. The distribution parameterized by a descriptor $w$ would be $p(\cdot;w)$ is common and shared amongst individuals (cite?.Following the aforementioned example statement, this would imply $p_{\theta_A}(\theta_A) = p(\theta_A;\text{"rarely never"})$.

## Higher Order Statements

A first order probability statement would consist of a descriptor that only uses one word such as "never". It is also possible to have higher order probability statements such as "Event A rarely never happens", where the descriptor is a combination of two or more descriptors. The following section would focus on deriving the form of $p(\theta_A;w_1,w_2)$ from $p(\theta_A;w_1)$ and $p(\theta_A;w_2)$.

Given a general statement "Event A $w_1$ $w_2$ happens", we are interested in $p_{\theta_A}(\theta_A)$, the distribution of the parameter in the Bernoulli distribution of event A. We define a new Bernoulli event, event B, as the event that $w_2$ is a valid descriptor of event A, parameterized by $\theta_B$, which is distributed according to $p_{\theta_B}(\theta_B;w_1)$.

$$p_{B|\theta_B}(B|\theta_B) = \begin{cases} \theta_B & B = 1, p_{\theta_A}(\theta_A) = p(\theta_A;w_2) \\ 1-\theta_B & \text{else} \end{cases} \tag{3}$$

And by applying equation 2, we would get that p(B=1) is the expected value of the distribution parmeterized by $w_1$, and by linearity of expectation, $p(B=0)$ would be:

$$\begin{aligned} p(B=0) &= E_{p_{\theta_B}}[1-\theta_B] \\ &= 1 - E_{p_{\theta_B}(\cdot;w_1)}[\theta_B] \end{aligned} \tag{4}$$

**Negated distribution.** When $B=1$, $p_{\theta_A}(\theta_A)$ is accurately described by $w_2$. However, when $B=0$, all we know is that $p_{\theta_A}(\theta_A) \neq p(\theta_A;w_2)$ and a naive way to understand this is to sample a distribution $q$ with probability $p_q(q)$ from the space of all possible distributions $Q$ without the distribution parameterized by $w_2$. We would denote $Q \setminus p(\cdot;w_2)$ as $Q'$, and $p_q()$ be a distirbution over $Q'$. So under the scenario $B=0$ by bayes rule:

$$\begin{aligned} p_{\theta_A}(\theta_A = n|B=0) &= \sum_{q\in Q'} p_q(q) q(\theta_A = n) \\ &= E_{p_q}[q(\theta_A = n)] \end{aligned} \tag{5}$$

It is straightforward to see that $p_{\theta_A}(\theta_A = n|B=0)$ is a distribution as $q(\theta_A = n)$ is always positive and:

$$\begin{aligned} \int_{n=0}^1 p_{\theta_A}(\theta_A = n|B=0)dn &= \sum_{q\in Q'} p_q(q) \int_{n=0}^1 q(\theta_A = n)dn \\ &= \sum_{q\in Q'} p_q(q) = 1 \end{aligned} \tag{6}$$

This shows that $p_{\theta_A}(\theta_A|B=0)$, or the distribution of "not $w_2$" is a distribution that could be modelled, denoted as $p_{\theta_A}(\theta_A;n(w_2))$. However, there are still problems as we do not know whether humans sample from or have access to $Q'$, nor do we know $p_q$.

Taking inspiration from the semantic meanings of probabilistic statements, we proposed an optimization problem to find a distribution $Q$ which aligns with those semantic meanings. This is best illustrated by example. For instance, the word "never" theoretically conveys a deterministic distribution where all the weight is concentrated on the likelihood 0. Equation 7 considers the case where event A "never" happens, and $\theta_A$ parameterizes the Bernoulli event A.

$$p_{\theta_A|never}(\theta_A|never) = \begin{cases} 1 & \text{if } \theta_A = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Next, equation 8 considers the case where event A does **not** "never" happen.

$$p_{\theta_A|\text{not never}}(\theta_A|\text{not never}) = \begin{cases} 0 & \text{if } \theta_A = 0 \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

In this case, the distribution encoding "not never" exactly maximizes KL divergence ($P_\theta(\theta=0)=1$, and $Q_\theta(\theta=0)=0$, so the term in divergence $P_\theta(\theta=0)log(\frac{P_\theta(\theta=0)}{Q_\theta(\theta=0)}) \to \infty$) with the distribution "never" while also maximizing entropy. The divergence maximization follows logically; the negated word distribution ”not never” should be as far as possible from ”never”, and the entropy maximization further constrains the optimization problem to converge on a distribution with as little other information as possible, because semantically the phrase ”not never” does not tell us anything more than the idea that $\theta_A \neq = 0$.

Concretely, we posed the problem of finding the negated distribution as:

$$\hat{Q} = \text{argmax}_Q D_{KL}(P||Q) + \alpha H(Q) \tag{9}$$

where $\alpha$ is a constant associated with the information minimization term. $\alpha$ can be interpreted as a weighting of the importance of information content in the negated distribution.

**Resulting distribution.** With a negated distribution for $p_{\theta_A}(\theta_A; w_2)$, we can revisit the statement ”Event A $w_1$ $w_2$ happens” and event B (Equation 4), to obtain the final expression for $p_{\theta_A}(\theta_A; w_1, w_2)$, which ends up a linear combination of the $w_2$ distribution and $n(w_2)$ distribution.

$$\begin{aligned} p_{\theta_A}(\theta_A; w_1, w_2) = {} & p_{\theta_A}(\theta_A; w_2)p(B=1) + p_{\theta_A}(\theta_A; n(w_2)) \\ & p(B=0) \\ & = p_{\theta_A}(\theta_A; w_2)E_{p_{\theta_B}(\cdot; w_1)}[\theta_B] + p_{\theta_A}(\theta_A; n(w_2)) \\ & (1 - E_{p_{\theta_B}(\cdot; w_1)}[\theta_B]) \end{aligned} \tag{10}$$

## Experiment

### Psychophysical Experiment

In order to obtain samples from the distributions parameterized by simple and high order descriptors, we conducted a psychophysical experiment on participants using Amazon Mechanical Turk.

The experiment consisted of a fictional urn with 100 balls that are either red or blue, and a probabilistic statement describing the state of the urn. We chose a consistent experimental setup over examples as Brun (1988) has shown that ones judgement on probability could be context-based. Example statements include ”Red balls are always drawn from the urn” or ”Red balls are never drawn from the urn”. The participants were then asked to estimate the number of red balls inside the urn by adjusting a slider ranging from 0 to 100. We interpreted their answers as their predicted estimates of a Bernoulli parameter governing the identity of a ball drawn at random from the urn.

**Participants** 50 participants with IP adresses in the US were recruited on Amazon Mechanical Turk and compensated 1.5 USD for their responses.

**Questions Layout** Participants were presented with probabilistic statements in a randomized order. The statements were of the general format ”Red balls are [$word_1$] [$word_2$] drawn from the urn”, where $word_1$ could be any of ”always”, ”often”, ”usually”, ”occasionally”, ”rarely” or ”never”. $word_2$ was either omitted from the phrase (referred to as ”single word questions”) or was a word from the same set (we chose not to query participants with phrases that had the same word repeated). This resulted in 36 individual questions.

### Subject Filtering

A key assumption in our experiment is that all participants were sampling from an internal distribution governed by the probabilistic word. Further, we treated each participants’ response as a sample from the same collective distribution. To ensure this assumption was valid, selected for further analysis only participants whose responses were generally correlated to each other. Concretely, we decomposed response profiles of participants using principle components analysis into 2 dimensions and then ran a KMeans clustering search using $1-r$ as a distance metric between participants. The data naturally separated into 2 clusters (figure 1 A), one of which consisted of a group of participants whose responses were highly correlated, and the other a group of participants whose responses were highly uncorrelated to each other. Only the participants which belonged to the former cluster were used for later analysis ($n = 24$).

## Simple expressions

Since all of our participants are sampling from the same distribution that is specified by the word of interest, this allows us to treat each participant’s response as a sample from the same distribution. Given these 21 samples of $p_{\theta_A}(\theta_A; w)$, the results are shown in figure 1B.

### Results

Figure 1B shows the distributions for each word, and figure 1 shows the statistics for these distributions. As expected never and always has extreme mean values of 7.0 and 95.3 and other distributions have less extreme means. However, the variances are much more surprising - we expected ”never” and ”always” to have a relative small variance as there is little to no ambiguity, and variance be high in ambiguous terms like ”occasionally”; instead, we saw that ”never” has a surprisingly high variance of 17.0, approaching ”occasionally”, which has the highest variance of 20.4. The word with lowest variance is ”rarely” with 7.9 variance, which shows that rarely is a rather unambiguous term amongst participants.

Table 1: Mean/variance of single word empirical distributions

| | Never | Rarely | Occasionally | Usually | Often | Always |
|---|---|---|---|---|---|---|
| *Mean* | 7.0 | 11.0 | 48.4 | 84.3 | 81.4 | 95.3 |
| *Variance* | 17.0 | 7.9 | 20.4 | 10.1 | 8.7 | 8.2 |

### Analysis

We chose to fit beta distributions to parameterize the empirical results of single word questions, as these are naturally used to model priors of Bernoulli events and can be defined

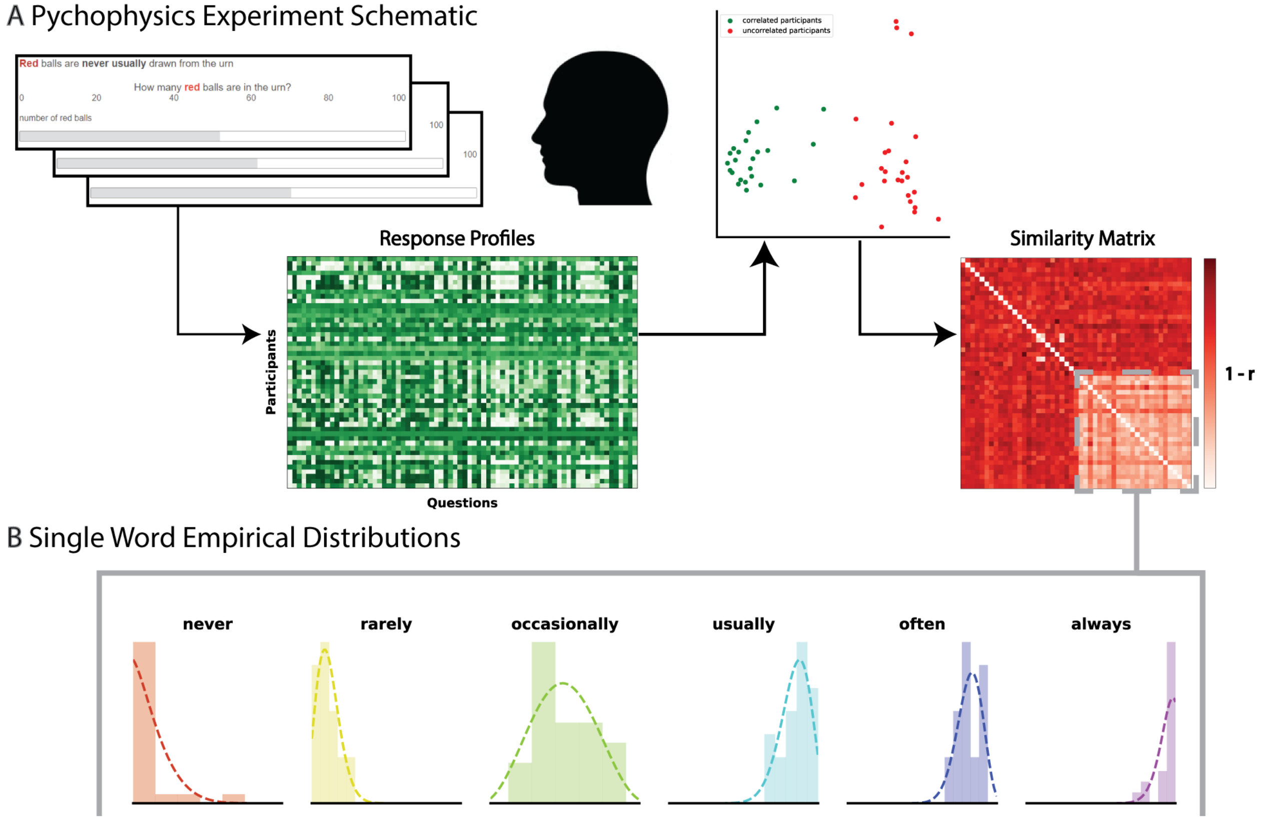
A Pychophysics Experiment Schematic
Red balls are never usually drawn from the urn
How many red balls are in the urn?
0
20
40
60
80
100
number of red balls
100
100
correlated participants
uncorrelated participants
Response Profiles
Participants
Questions
Similarity Matrix
1 - r
B Single Word Empirical Distributions
never
rarely
occasionally
usually
often
always

Figure 1: **Data Collection** A) Psychophysics experiment schematic: We queried participants on Amazon Mechanical Turk with 74 probabilistic statements ($n = 50$ participants). Participants' response profiles ([50,74] matrix) were decomposed into 2 dimensions using Principle Components Analysis, and were then clustered with KMeans ($k = 2$). Participants which were categorized into the cluster with a higher correlation were used for further analysis, and the rest were discarded (n=24). B) Fitted beta distributions (dashed line) plotted over empirical distributions (histogram). In general, the expectations of the fitted distributions matched accepted definitions of our query words.

over a boundary consistent with our construction of the problem. The continuous parameterized distributions also allow for realistic extensions of experimental data given our limited dataset. In this case, the beta distributions represent proposed internal distributions that participants sampled from, and the values of the distribution represent Bernoulli likelihoods for the event that a red ball is drawn from the urn.

Figure 1B shows the shape of the beta distributions fit to the corresponding data against empirical results. To assess the goodness of fit, we report Pearson correlations of the fitted beta distributions in table 2.

Table 2: Results of Pearson's Correlation test between beta distribution model and data

| | Never | Rarely | Occasionally | Usually | Often | Always |
|---|---|---|---|---|---|---|
| $R^2$ | 0.96 | 0.90 | 0.99 | 0.94 | 0.76 | 0.83 |
| $p$ | 2e-4*** | 8e-13*** | 3e-3*** | 7e-7*** | 3e-7*** | 9e-6*** |

## High Order expressions

### Negated Distributions

For each single descriptor model with word w, we computed the negated distribution corresponding to $n(w)$. To do so, we used stochastic gradient descent with a learning rate of 0.01 to optimize a discrete vector $v \in \Re^{101}$, with $v[k]$ representing $p(k;n(w))$. The discrete representation of the negated distribution was sufficient for the calculation of theoretical distributions governing two-word phrases.

Following the form described in equation 9, we chose $\alpha = 1$. The final form of the discrete optimization is described in equation 11.

$$\hat{v} = \text{argmax}_v \sum_{k=0}^{100} p(k) log \frac{p(k)}{v[k]} + \sum_{k=0}^{100} v[k] log v[k] \qquad (11)$$

The resulting interpolated negated distributions are shown in figure 2A with the original distributions as a dotted line.

## High Order Expression

After calculating the negated distributions, we can obtain a discrete distribution of high order expressions by using a discrete version of Equation 10 (Equation 12).

$$p_{\theta_A}[\theta_A = k; w_1, w_2] = p(k; w_2)E[\theta_b] + p[k; n(w_2)](1 - E[\theta_b]) \tag{12}$$

Note that $E[\theta_b]$ is just the mean of $p(\cdot; w_1)$ which is the distribution parametrized by $w_1$. With values for $p[k; w_1, w_2]$, we can run a correlation test between the theoretical distribution with the empirical data to test the validity of our model.

## Results

Table 3 show the $R^2$ values of the test. Considering a sample size of 24, the correlation are strong, especially for probabilities with extreme means such as ”never” and ”always”, as the first word. This is likely due to the fact that this means the resulting distribution mostly echos the original first order distribution or the negated distribution, which indicates that our negated distribution does indeed capture a human’s interpretation of ’not a distribution’ according to Equation 6.

For complex statements that involve more ambiguous words such occasionally, we see that there are some innate biases in humans. For example ”rarely occasionally” in Figure 2B shows a peak in the empirical data that is not accounted for in the theoretical distribution. This demonstrates a innate bias towards 0 when $w_1$ and $w_2$ both have means close to 0, possibly related to the idea of litotes.

To this end, several of our proposed second-order phrase distributions appeared to be shifted versions of the empirical distributions (for example, ”never rarely”). This indicates that our psychophysics data did not consistently represent single word distributions in the manner we had anticipated. However, for a majority of the two-word combinations, we observed strikingly similar empirical and predicted theoretical distributions.

Another trend was that our modeled distributions more closely matched empirical distribution as the modifier adverb’s (the first word) distribution expectation was closer to 1 (figure 3B). We can conclude that our models suffer when the semantic meaning is further obscured by close-to-zero mean modifiers.

Table 3: $R^2$ values of Pearson’s Correlation test between model and empirical

| | Never | Rarely | Occasionally | Usually | Often | Always |
|---|---|---|---|---|---|---|
| Never | -0.13 | 0.81 | 0.29 | -0.09 | 0.08 | |
| Rarely | -0.69 | | 0.31 | 0.29 | 0.74 | 0.01 |
| Occasionally | 0.99 | 0.40 | | -0.46 | 0.86 | 0.71 |
| Usually | 0.84 | 0.94 | 0.58 | | 0.28 | 0.94 |
| Often | 0.98 | 0.94 | 0.65 | 0.01 | | 0.97 |
| Always | 0.78 | 1.00 | -0.06 | 0.68 | -0.68 | |

## Conclusions and Future Work

We first demonstrated that the process of understanding a first order probabilistic statement with frequency adverbs can be adequately modelled by sampling from a beta distribution. Then, we derived a semantically consistent idea of a negated distribution by finding a distribution that maximizes KL divergence and entropy.

Our work has also demonstrated a methodology for generating theoretical distributions of second order probabilistic descriptors using only distributions from first order descriptors. This allows us to arbitrarily compute the distributions of higher order descriptors, which offers a unique insight regarding probabilistic statements - that they can be derived from linear combinations of the first order distributions and negated 1st order distributions.

These higher-order statements reveal uncertainty about the probabilities themselves, unravelling an extra layer of complexity to the NLP problem of interpreting human use of probabilistic phrases. Furthermore, we observed strong correlations between the computed distributions governing second-order statements and the empirical data found through psychophysics experimentation, suggesting that the internal process of sampling which humans use to evaluate probabilistic phrases may be similar to the hierarchical procedure we developed.

Ultimately, our novel interpretation of the meaning of probabilistic adverbs, namely as prior distributions governing a $\theta$ which parameterizes Bernoulli distributions, lends insight into human language processing. Future work will focus on improvements to our negative-word distribution formulation in order to more closely match samples of empirical data.

## References

Herbstritt, M., & Franke, M. (2019). Complex probability expressions higher-order uncertainty: Compositional semantics, probabilistic pragmatics experimental data. *Cognition*, *186*, 50 - 71. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0010027` doi: https://doi.org/10.1016/j.cognition.2018.11.013

Moss, S. (2015, March). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*(5), 1–81. doi: 10.3765/sp.8.5

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few many. *Linguistics Vanguard*(1), 20160072. Retrieved from `https://www.degruyter.com/view/journals/lingvan/open-issu` doi: https://doi.org/10.1515/lingvan-2016-0072

Tessler, M. H., & Goodman, N. D. (2018). *The language of generalization.*

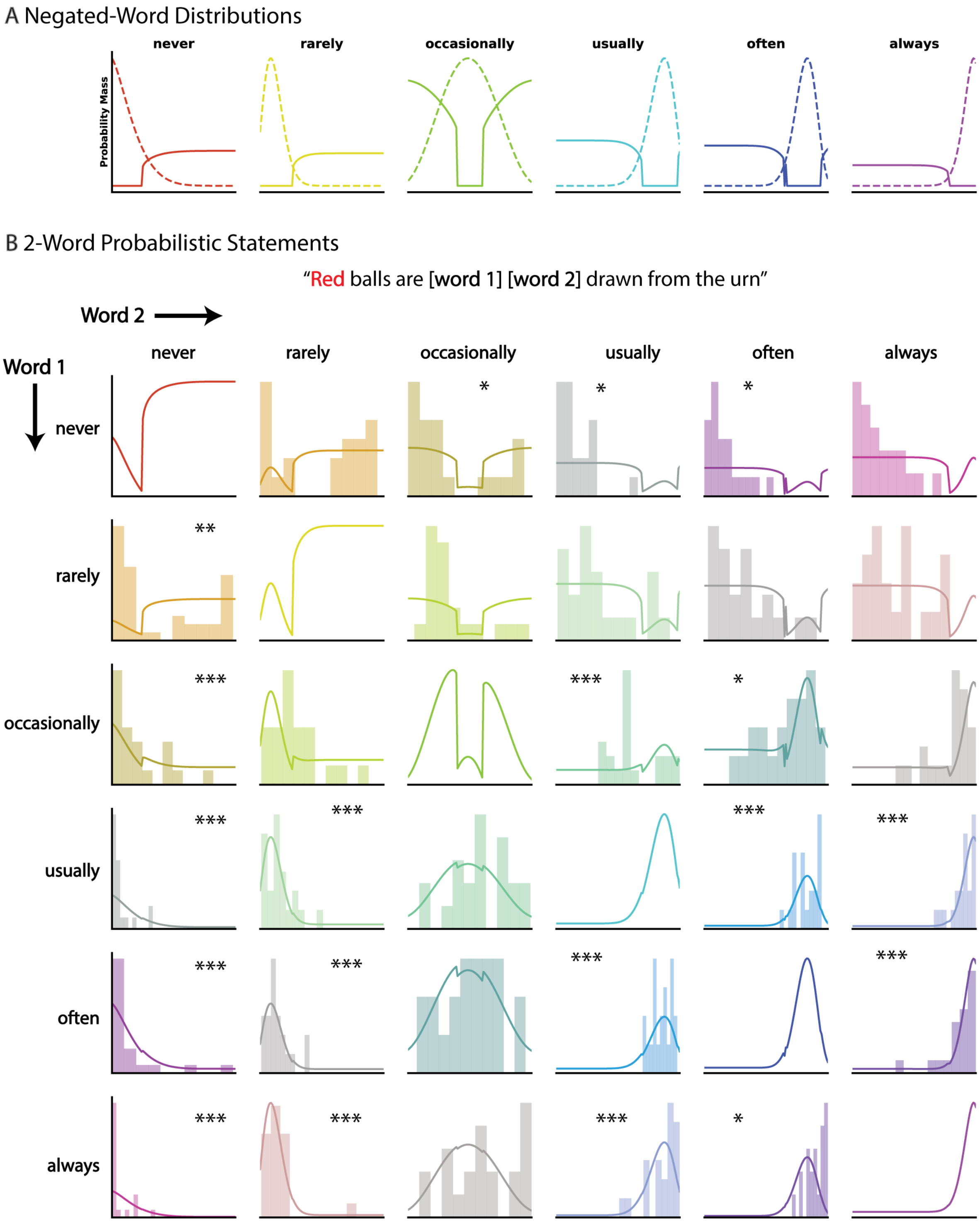

Figure 2: **2-Word Statement Models** A) negated distributions for each word (solid line) compared to parameterized beta distribution fit to data (dashed line). B) Empirical data (histogram) compared to theoretical 2-word distribution. No data was recorded for phrases with repeated words. Rows correspond to the first word and columns correspond to the second word (* indicate Pearson correlation p-value; $* \leftarrow p \leq 0.01$, $** \leftarrow p \leq 0.001$, and $*** \leftarrow p \leq 0.0001$).