# A Bayesian Model of Grounded Color Semantics

**Brian McMahan**
Rutgers University
brian.mcmahan@rutgers.edu

**Matthew Stone**
Rutgers University
matthew.stone@rutgers.edu

## Abstract

Natural language meanings allow speakers to encode important real-world distinctions, but corpora of grounded language use also reveal that speakers categorize the world in different ways and describe situations with different terminology. To learn meanings from data, we therefore need to link underlying representations of meaning to models of speaker judgment and speaker choice. This paper describes a new approach to this problem: we model variability through uncertainty in categorization boundaries and distributions over preferred vocabulary. We apply the approach to a large data set of color descriptions, where statistical evaluation documents its accuracy. The results are available as a Lexicon of Uncertain Color Standards (LUX), which supports future efforts in grounded language understanding and generation by probabilistically mapping 829 English color descriptions to potentially context-sensitive regions in HSV color space.

## 1 Introduction

To ground natural language semantics in real-world data at large scale requires researchers to confront the vocabulary problem (Furnas et al., 1987). Much of what people say falls in a long tail of increasingly infrequent and specialized items. Moreover, the choice of how to categorize and describe real-world data varies across people. We can't account for this complexity by deriving one definitive mapping between words and the world.

We see this complexity already in free text descriptions of color patches. English has fewer than a dozen basic color words (Berlin, 1991), but people's descriptions of colors are much more variable than this would suggest. Measured on the corpus described in Section 4.1, there's an average of 3.845 bits of information in a color description *given* the color it describes—comparable to rolling a 14-sided die. Figure 1 summarizes the data and plots the entropy of descriptions encountered within small bins of color space. The bins are aggregated over the Saturation and Value dimensions and indexed on the x-axis by the Hue dimension. There's little reason to think that this variability conceals consistent meanings. In formal semantics, one of the hallmarks of vague language is that speakers can make it more precise in alternative, incompatible ways (Barker, 2002). We see this in practice as well, for example with the image of Figure 2, where subjects com-

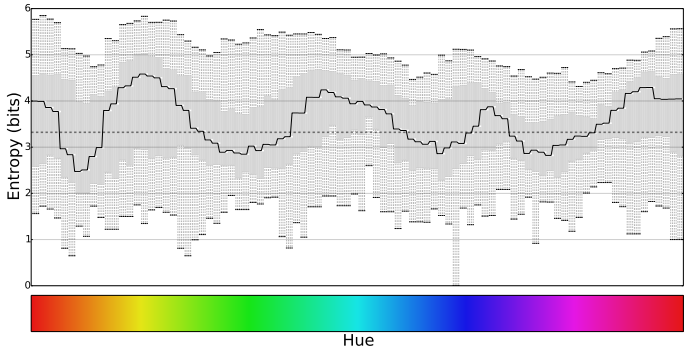


Figure 1: A visualization of the variability of the descriptions used to name colors within small bins of color space. For each Hue value, the entropy values for each bin along the Saturation and Value dimensions are grouped and plotted as box plots. The dotted line corresponds to a random choice out of fourteen items and to the perplexity of a histogram model trained on the corpus.



Transactions of the Association for Computational Linguistics, vol. 3, pp. 103–115, 2015. Action Editor: Lillian Lee.
Submission batch: 11/2014; Published 2/2015. 

Figure 2: Image by flickr user Joanne Bacon (jlbacon) from the data set of Young et al. (2014), whose subjects describe these dogs as *a brown dog and a tan one* or *a tan dog and a white one*.

prehensibly describe either of two dogs as *the tan one*. Systems that robustly understand or generate descriptions of colors in situated dialogue need models of meaning that capture this variability.

This paper makes two key contributions towards this challenge. First, we present a methodology to infer a corpus-based model of meaning that accounts for possible differences in word usage across different speakers. As we explain in Section 2, our approach differs from the typical perspective in grounded semantics (Tellex et al., 2011a; Matuszek et al., 2012; Krishnamurthy and Kollar, 2013), where a meaning is reduced to a single classifier that collapses patterns of variation. Instead, our model allows for variability in meaning by positing uncertainty in classification boundaries that can get resolved when a speaker chooses to use a word on a specific occasion. We explain the model and its theoretical rationale in Section 3.

Second, we develop and release a Lexicon of Uncertain Color Standards (LUX) by applying our methodology to color descriptions. LUX is an interpretation of 829 distinct English color descriptions as distributions over regions of the Hue–Saturation–Value color space that describe their possible meanings. As we describe in Section 4, the model is trained by machine learning methods from a subset of Randall Munroe's 2010 publicly-available corpus of 3.4 million crowdsourced free-text descriptions of color patches (Munroe, 2010). Data, models and visualization software are available at `http://mcmahan.io/lux/`.

Statistical evaluation of our model against two alternative approaches documents its effectiveness. The model makes better quantitative predictions than a brute-force memorization model; it seems to generalize to unseen data in more meaningful ways. At the same time, our meanings work as well as special-purpose models to explain speaker choice, even though our model supports diverse other reasoning. See Section 5.

We see color as the first of many applications of our methodology, and are optimistic about learning vague meanings for other continuous domains as quantity, space, and time. At the same time, the methodology opens up new prospects for research on negotiating meaning interactively (Larsson, 2013) with principled representations and with broad coverage. In fact, many practical situated dialogue systems already identify unfamiliar objects by color. We expect that LUX will provide a broadly useful resource to extend the range of descriptions such systems can generate and understand.

## 2 Related Work

Grounded semantics is the task of mapping representations of linguistic meaning to the physical world, whether by perceptual mechanisms (Harnad, 1990) or with the assistance of social interaction (DeVault et al., 2006). In this paper, we are particularly concerned with grounding the meanings of primitive vocabulary. However, the ultimate test of grounded semantics—whether it is understanding commands (Winograd, 1970; Tellex et al., 2011b), describing states of the world (Chen and Mooney, 2008), or identifying objects (Matuszek et al., 2012; Krishnamurthy and Kollar, 2013; Dawson et al., 2013)—is the ability to interpret or generate utterances using lexical and compositional semantics so as to evoke appropriate real-world referents. Grounded semantics therefore involves more than just quantifying the associations between words and perceptual representations, as Chuang et al. (2008) and Heer and Stone (2012) do for color. Grounded semantics involves interpreting semantic primitives in terms of composable categories that let systems discriminate between cases where a word applies and cases where the word does not apply. (Our evaluation compares models of grounded semantics to more direct models of word–world associations.)

Previous research has modeled these categories as

regions of suitable perceptual feature spaces. Researchers have explored explicit spaces of high-level perceptual attributes (Farhadi et al., 2009; Silberer et al., 2013), approximations to such spaces (Matuszek et al., 2012), or low-level feature spaces such as Bag of Visual Words (Bruni et al., 2012) or Histogram of Gradients (Krishnamurthy and Kollar, 2013). We specifically follow Gärdenfors (2000) and Jäger (2010) in assuming that color categories are convex regions in an underlying color space, and are not just determined by prototypical color values, such as in Andreas and Klein (2014).

However, unlike previous grounded semantics, we do not assume that words name categories unequivocally. Speakers may vary in how they interpret a word, so we treat the link between words and categories probabilistically. The difference makes training our model more indirect than previous approaches to grounded meaning. In particular, our model introduces a new layer of uncertainty that describes what category the speaker uses.

Similar kinds of uncertainty can be found in Bayesian models of speaker strategy, such as that of Smith et al. (2013). However, this research has assumed that speakers aim to be as informative as possible. We have no evidence that our speakers do that. We assume only that speakers' utterances are reliable and mirror prevailing usage.

Prior work by cognitive scientists has studied color terms extensively, but focused on basic ones—monolexemic, top-level color words with general application and high frequency in a language (Kay et al., 2009; Lammens, 1994). These color categories seem to shape people's expectations and memory for colors (Persaud and Hemmer, 2014), and patterns of color naming can therefore enhance software for helping people organize and interact with color (Chuang et al., 2008; Heer and Stone, 2012). Moreover, crosslinguistic evidence suggests that the human perceptual system places strong biases on the meanings of the basic color terms (Regier et al., 2005), perhaps because basic terms must partition the perceptual space in an efficient way (Regier et al., 2007). We depart from research on basic color naming in considering a much wider range of terms, much like Andreas and Klein (2014). We consider subordinate, non-basic terms like *beige* or *lavender*; modified colors like *light blue* or *bright green*; and named subcategories like *olive green*, *navy blue* or *brick red*.

In order to use semantic primitives for understanding, it's necessary to combine them into an integrated sentence-level representation: this is the problem of semantic parsing. Semantic parsers can be built by hand (Winograd, 1970), induced through inductive logic programming (Zelle and Mooney, 1996), or treated as a structured classification problem (Zettlemoyer and Collins, 2005). Once a suitable logical form is derived, interpretation typically involves a recursive process of finding referents that fit lexical categories and relationships (Mavridis and Roy, 2006; Tellex et al., 2011a). While this paper does not explicitly address how our meanings might be used in conjunction with such techniques, we see no fundamental obstacle to doing so—for example, by resolving references probabilistically and marginalizing over uncertainty in meaning.

## 3 Using Vague Color Terms: A Model

Our model involves two significant innovations over previous approaches to grounded meaning. The first is to capture the vagueness and flexibility of grounded meaning with semantic representations that treat *meaning* as uncertain. We represent the semantics of a color description with a *distribution* over color categories, which weights possible meanings by the relative likelihood of a speaker using this meaning on any particular occasion. For example, speakers might associate *yellowish green* with a range of possible meanings, differing in how far the color category extends into green hues. By representing uncertainty about meaning, our model makes room to capture variability in language use. For example, it implicitly quantifies how likely speakers are to use words differently, as with the two interpretations of *tan* in Figure 2.

Our second contribution is our simple model of the relationship between semantics and pragmatics. We assume that speakers' choices mirror established patterns. In particular, the model learns a measure of *availability* for each color term that tracks how frequently speakers tend to use it when it is applicable. For example, although the expressions *yellowish green* and *chartreuse* are associated with very similar color categories, people say *yellowish green*

much more often: it has a higher availability. Empirically, we find few terms with high availability and a long tail of terms with lower availabilities. We assume speakers simply sample applicable terms from this distribution, which predicts the long tail of observed responses.

Mathematically, we develop our approach through the rational analysis methodology for explaining human behavior proposed by Anderson (1991), along with methodological insights from the linguistics and philosophy of vagueness. In the remainder of this section, we explain the theoretical antecedents in perceptual science, linguistics and cognitive modeling that inform our approach.

### 3.1 Color Categories

Color can be defined as sensations by which the perceptual system tracks the diffuse reflectance of objects, despite variability, uncertainty and ambiguity in the visual input. Red, green, and blue cones in the retina allow the visual system to coarsely estimate frequency bands in the spectrum of incoming light. Cameras and screens that use the red–green–blue (RGB) color space are designed roughly to correspond to these responses. However, colors in the visual system summarize spectral profiles rather than mere wavelengths of light. For example, we see colors like cyan (green plus blue without red), magenta (blue plus red without green) and yellow (red plus green without blue) as intermediate saturated colors between the familiar primaries. This naturally leads to a wheel of hues describing the relative prominence of different spectral components along a continuum. Fairchild (2013) provides an overview of color appearance.

To capture this variation, we'll work in the simple hue–saturation–value (HSV) color space that's common in computer graphics and color picker user interfaces (Hughes et al., 2013) and implemented in python's native colorsys package. This coordinate system represents colors with three distinct qualitative dimensions: *Hue* (H) represents changes in tint around a color wheel, *Saturation* (S) represents the relative proportion of color versus gray, and *Value* (V) represents the location on the white–black continuum. We will associate color categories with rectangular box-shaped regions in HSV space. More sophisticated color spaces have been developed to describe the psychophysics of color more precisely, but they depend on the photometric illumination and other aspects of the viewing context that were not controlled in the collection of the data we are using (Fairchild, 2013).

### 3.2 Semantic Representation

Our assumption is that color terms are associated probabilistically with color categories. We illustrate the idea for the color label *yellowish-green* through the plot in Figure 3. The plot shows variation in use of the term across the *Hue* dimension: the bar graph is a scaled histogram of the responses in the data we use. There is a range of colors where people use *yellowish green* often, surrounded by borderline cases where it becomes increasingly infrequent.

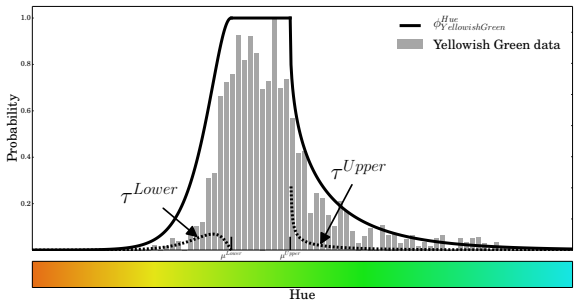


Figure 3: The LUX model for "yellowish green" on the *Hue* axis plotted against the scaled histogram of the responses in the data. The $\phi$ curve represents the likelihood of "yellowish green" for different *Hue* values. The $\tau$ curves represent possible boundaries.

We represent this variability by assuming that the boundaries that delimit the color are uncertain. In any utterance, *yellowish green* fits only those *Hue* values that are above a minimum threshold $\tau^{Lower}$ and below a maximum threshold $\tau^{Upper}$. However, it is uncertain which thresholds a speaker will use. The model describes this variability with probability density functions. They are shown for *yellowish green* in Figure 3 as the $\tau$ distributions. The figure shows that there is a central range of hues, between the $\tau$ distributions, that is definitely *yellowish green*. The $\tau$ distributions peak at the most likely boundaries for *yellowish green*, encompassing a broad region that's frequently called *yellowish green*. Further away, threshold values and *yellowish green* utterances alike become rapidly less likely.

Our representation is motivated by Barker (2002) and Lassiter (2009), who show how sets of possible thresholds[1] can account for many of our intuitions about the use of vague language. Their analysis invites us to capture semantic variability through two geometric constructs. First, there is a *certain* interval, parameterized by two points, $\mu^{Lower}$ and $\mu^{Upper}$, within which a color description definitely applies. Outside this interval are regions of borderline cases, delimited by probabilistically-varying thresholds $\tau^{Lower}$ and $\tau^{Upper}$, where the color description sometimes applies. We represent the position of the threshold with a $\Gamma(\alpha, \beta)$ distribution, a standard statistical tool to model processes that start, continue indefinitely, and stop, like waiting times.[2] We can determine a likelihood that a description fits a color by marginalizing over the thresholds: this gives the black curve visualized in Figure 3. As we describe in Section 3.3, we can use this to account for the graded responses from subjects that we observe near color boundaries.

We summarize with a formal definition of our semantic representation. Let $X$ be the 3D space of HSV colors and let $x \in X$ be a measured color value. Each color label $k$ has definite boundaries, $\mu^{Lower}$ and $\mu^{Upper}$ in $X$, delimiting a box of HSV color space. Surrounding the definite region are regions of uncertainty: the set of possible boundaries beyond $\mu$. These are represented by probability distributions over lower and upper threshold values in each dimension. We'll represent these thresholds by $\tau_k^{j,d}$ where $k \in K$ indexes the color label, $j \in \{Lower/L, Upper/U\}$ indexes the boundary, and $d \in \{H, S, V\}$ indexes color components. We assume the thresholds are distributed as follows:

$$\tau_k^{Lower,d} \sim \mu_k^{Lower,d} - \Gamma(\alpha_k^{Lower,d}, \beta_k^{Lower,d})$$
$$\tau_k^{Upper,d} \sim \mu_k^{Upper,d} + \Gamma(\alpha_k^{Upper,d}, \beta_k^{Upper,d}) \quad (1)$$

The meaning of a color term is thus a "blurry box". The distribution lets us determine the probability of a point $x$ falling into the color category $k$ as in Eq. 2. We also use the compact notation in Eq. 3.

$$P(\tau_k^{Lower,H} < x^H < \tau_k^{Upper,H}) \times$$
$$P(\tau_k^{Lower,S} < x^S < \tau_k^{Upper,S}) \times$$
$$P(\tau_k^{Lower,V} < x^V < \tau_k^{Upper,V}) \quad (2)$$
$$= \prod_d P(\tau_k^{L,d} < x_i^d < \tau_k^{U,d}) \quad (3)$$

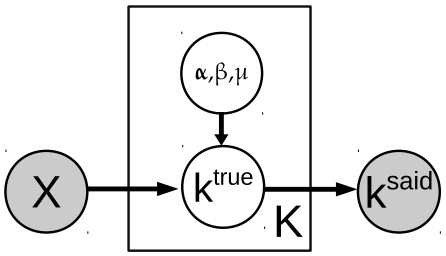


Figure 4: The Rational Observer observes a color patch, $x$. The applicability of each label ($k^{true}$) is based upon the label parameters ($\alpha, \beta, \mu$) and $x$. The label ($k^{said}$) is sampled proportional to the applicability and a background weight: how often a label is said when it applies.

### 3.3 Rational Observer Model

Our goal is to learn probabilistic representations of the meanings of color terms from subjects' responses. To do this, we need not only a framework for representing colors but also a model of how subjects choose color terms. Inspired by rational analysis (Anderson, 1991), we assume that speakers' choices match their communicative goals and their semantic knowledge. We leverage this assumption to derive a Bayes Rational Observer model linking semantics to observed color descriptions.

The graphical model in Figure 4 formalizes our approach. We start from an observed color patch, $x$. The Rational Observer uses the $\tau$-distributions for each color description $k$ to determine the likelihood that the speaker judges $k$ applicable. As defined in Eq. 3, the likelihood is the subset of possible boundaries which contain the target color value. Normally, many descriptions will be applicable. Which the speaker chooses depends further on the *availability* of the label—a background measure of how frequently a label is chosen when it's applicable. Intuitively, availability creates a bias for easy descriptions, capturing how natural or ordinary a descrip-

[1] We treat the terms "boundary", "threshold", and "standard" to be synonymous, but useful in different contexts.

[2] $\Gamma$ distributions rise quickly away from the origin point, then trail off from the peak in an open-ended exponential decay. One intuition for applying them in this case is Graff Fara's (2000) suggestion that a particular categorization decision involves waiting to find a natural break among salient colors. However, we choose them for mathematical convenience rather than psychological or linguistic considerations.

tion is in language use, how easily it springs to mind or how easily it is understood.

We formalize this as a generative model. As we explain in Section 4, we infer the parameters from our data. In Eq. 4, we consider the conditional distribution of a subject observing a color patch given HSV value $x$ and labeling it $k$:

$$P(k^{said}, k^{true}|x) = P(k^{said}|k^{true})P(k^{true}|x) \quad (4)$$

In this equation, $k^{said}$ is the event that the subject responds to $x$ with label $k$ and $k^{true}$ is the event that the subject judges $k$ true of the HSV value $x$. The two factors of Eq. 4 are respectively the *availability* and *applicability* of the color label.

**Availability**: The prior $P(k^{said}|k^{true})$ quantifies the rate at which label $k$ is used when it applies. We refer to this quantity as the *availability* and denote it as $\alpha_k$. Availability captures the observed bias for frequent color terms. When multiple color labels fit a color value, those with higher availability will be used more often, but those with lower availability will still get used. This effect is partially responsible for the long tail of subjects' responses.

**Applicability**: The second factor, $P(k^{true}|x)$, is the probability that $k$ is true of, or applies to, the color value $x$. We calculate the applicability by marginalizing over all possible thresholds as in Eq. 3. In other words, we calculate the probability mass of the boundaries which allow for this description to apply. We treat each applicability judgment as independent of others. This implies that the relative frequency at which we see a color description used is directly proportional to the proportion of boundaries which license it.

For clearer notation and parameter estimation, we track thresholds with a piecewise function $\phi_k^d(x^d)$ as in Eq. 5 and Figure 3.

$$\phi_k^d(x^d) = \begin{cases} P(x^d > \tau_k^{L,d}), & x^d \leq \mu_k^{L,d} \\ P(x^d < \tau_k^{U,d}), & x^d \geq \mu_k^{U,d} \\ 1, & otherwise \end{cases} \quad (5)$$

Finally, Eq. 6 rewrites Eq. 4 to make the applicability and availability explicit. The model treats this equation as the probability of success for a Bernoulli trial and the data as sampled from Categorical distributions formed by the set of $K$ Bernoulli random variables. This is discussed further in Section 4.2.

$$P(k^{said}, k^{true}|\mathbf{x}) = \alpha_k \prod_d \phi_k^d(x^d) \quad (6)$$

## 4 Learning Experiment

We worked with Randall Munroe's crowdsourced corpus of color judgments, and fit the model using the Metropolis-Hastings Markov Chain Monte Carlo, a Gaussian random walk optimization method. This form of approximate Bayesian inference is described in Section 4.2.

### 4.1 Munroe Color Corpus

In 2010, Munroe elicited descriptions of color patches over the web. His platform asked users for background information such as sex, colorblindness, and monitor type, then presented color patches and let the user freely name them. The setup didn't ensure that users see controlled colors or that users' responses are reliable, but the experiment collected over 3.4M items pairing RGB values with text descriptions. Munroe's methodology, data and results are published online (Munroe, 2010).[3]

Munroe summarizes his results with 954 idealized colors—RGB values that best exemplify high frequency color labels. In effect, Munroe's summary offers a prototype theory of color vocabulary, like that of Andreas and Klein (2014). An alternative theory, which we explore, is that variability in the applicability of labels is an important part of people's knowledge of color semantics. We compare the two theories explicitly in Section 5.

Our experiments focus on a subset of Munroe's data comprising 2,176,417 data points and 829 color descriptions, divided into a training set of 70%, a 5% development set, and a held-out test set of 25%. To minimize variability in language use, we selected data from users who self-report as non-colorblind English speakers. This accounts for 2.5M of Munroe's 3.4M items. To get our subset, we further restrict attention to labels used 100 times or more, to ensure that there's substantial evidence of each term's breadth of applicability. We hand curated the responses to correct some minor spelling variations involving a single-character

[3] http://blog.xkcd.com/2010/05/03/color-survey-results/

change ("yellow green" vs "yellow-green"; "fuchsia" vs "fuschia", "fushia", "fuchia", and "fucsia") and to remove high-frequency spam labels. We are left with 829 color labels that fit these restrictions. Finally, we used python's colorsys to convert from RGB to HSV, where we hypothesize color meanings can be represented more simply. We include these data sets with our release at `http://mcmahan.io/lux/` so our results can be replicated.

### 4.2 Fitting the Model Parameters

Optimization of the model's parameters is framed in a Bayesian framework and interpreted as maximizing the likelihood of the data given the parameters. We fit each label and each dimension independently. The data on each dimension is binned, as in Figure 3, so we have Binomial random variables for each bin. For each color label $k$, the probability of success is based on the model's parameters. Non-$k$ data in the bin are observations of failure. This gives Eq. 7:

$$P(n^d_{i,k}|n^d_i, Z^d_k, \phi_k) \sim Bin(n^d_i, Z^d_k \phi^d_k(i)) \qquad (7)$$

Here $n^d_i$ is the number of data points in bin $i$ on dimension $d$, $n^d_{i,k}$ is the number of data points for label $k$ in bin $i$ on dimension $d$, and $Z^d_k$ is a normalization constant, implicitly reflecting both the availability $\alpha_k$ and the distribution of responses of the term across other color dimensions. The optimization process is a parameter search method which uses as an objective function the probability of $n^d_{i,k}$ in Eq. 7 for all $d$,$i$, and $k$.

**Parameter Search**: We adopt a Bayesian coordinate descent which sequentially samples the certain region parameter, $\mu$, and the shape and rate parameters ($\alpha$ and $\beta$) of the $\Gamma$ distributions for all $d$ and $k$ independently. It also samples the estimated normalization constant, $Z^d_K$. More specifically, the sampling is done using Metropolis-Hastings Markov Chain Monte Carlo (Metropolis et al., 1953; Chib and Greenberg, 1995), which performs a Gaussian random walk on the parameters[4]. For each sample, the likelihood of the data, derived from the Binomial variables, is compared for the new and old set of parameters. The new parameters are accepted proportionally to the ratio of the two likelihoods. Multiple chains were run using 4 different bin sizes per dimension and monitored for convergence using the generalized Gelman-Rubin diagnostic method (Brooks and Gelman, 1998). This methodology leaves us not only with the Monte Carlo estimate of the expected value for each parameter, but also a sampling distribution that quantifies the uncertainty in the parameters themselves.

**Availability**: Availability is estimated as the ratio of the observed frequency of a label to its expected frequency given the parameters which define its distribution. The expected frequency, a marginalization of the color space for the $\phi$ function, is calculated using the midpoint integration approximation.

$$\begin{aligned} \alpha_k &= \frac{P(k^{said}, k^{true})}{P(k^{true})} \\ &= \frac{count(k)/N}{\int_x P(k^{true}|x)P(x)} \end{aligned} \qquad (8)$$

## 5 Model Evaluation

LUX explains Munroe's data via speakers' rational use of probabilistic meanings, represented as simple "blurry boxes". In this section, we assess the effectiveness of this explanation. We anticipate two arguments against our model: first, that the representation is too simple; second, that factoring speakers' choices through a model of meaning is too cumbersome. We rebut these arguments by providing metrics and results that suggest that LUX escapes these objections and captures almost all of the structure in subjects' responses.

### 5.1 Alternative Models

To test LUX's representations, we built a brute-force histogram model (HM) that discretizes HSV space and tracks frequency distributions of labels directly in each discretized bin. Similar histogram models have been developed by Chuang et al. (2008) and (Heer and Stone, 2012) to build interfaces for interacting with color that are informed by human categorization and naming. More precisely, our HM uses a linear interpolation method (Chen and Goodman, 1996) to combine three histograms of various

[4]We set the standard deviation of the sampling Gaussian to be 1 for each $\mu$ and 0.3 for each $\alpha$ and $\beta$ after finding experimentally that it led to effective parameter search (Gelman et al., 1996).

granularity.[5] This amounts to predicting responses by querying the training data. HM has the potential to expose whether LUX is missing important features of the distribution of color descriptions.

We also built a direct model of subjects' choices of color terms. Instead of appealing to the applicability and availability of a color label, it works with the observed frequency of a color label and a Gaussian model of the probability of a color value for each label, as in Eq. 9:

$$P(k^{said}, k^{true}|x) \propto P(x|k^{true})P(k^{said}, k^{true}) \quad (9)$$

This Gaussian model (GM) generalizes Munroe's pairing of labels with prototypical colors: $P(x|k^{true})$ is a Gaussian with diagonal covariance, so it associates each color term with a mean HSV value and with variances in each dimension that determine a label-specific distance metric. GM predicts speaker choice by weighting these distances probabilistically against the priors. GM completely sidesteps the need to model meaning categorically. It therefore has the potential to expose whether our assumptions about semantic representations and speaker choices hinder LUX's performance.

### 5.2 Evaluation Metrics

We evaluate the models using two classes of metrics on a held-out test set consisting of 25% of the corpus. The first type is based upon the posterior distribution over labels and the ranked position of subjects' actual labels of color values. The second type is based upon the log likelihood of the models, which quantifies model fit.

#### 5.2.1 Decision-Based Metrics

To answer how accurate a model's predictions are, we can locate subjects' responses in the weighted rankings computed by the models.

**The $\mathbf{TOP^K}$ Measures**: Each model provides a posterior distribution over the possible labels. The most likely label of this posterior is the maximum likelihood estimate (MLE). We track how often the MLE color label is what the user actually said as the $TOP^1$ measure. For the Histogram Model, the $TOP^1$ approximates the most frequent label observed in the data for a color value. We also measure how often the correct label appears in the first 5 and 10 most likely labels. These are denoted $TOP^5$ and $TOP^{10}$ respectively.

#### 5.2.2 Likelihood-Based Metrics

We can also measure how well a model explains speaker choice using the log likelihood of the labels given the model and the color values, denoted as $LLV(M)$. This is calculated using Eq. 10 across all $N$ data points in the held-out test set. $LLV(M)$ is used when computing perplexity and Aikake Information Criterion (AIC). We report all measures in bits.

$$\begin{aligned} LLV(M) &= \log_2 \ P_M(K^{true}, K^{said}|X) \\ &= \sum_i \log_2 \ P_M(k_i^{true}, k_i^{said}|x_i) \end{aligned} \quad (10)$$

A more general measure of model fit is the log likelihood of the color values and their labels jointly across the training set, $LL(V)$, given the model. It is defined and calculated analogously.

**Perplexity** Perplexity has been used in past research to measure the performance of statistical language models (Jelinek et al., 1977; Brown et al., 1992). Lower perplexity means that the model is less surprised by the data and so describes it more precisely. We use it here to measure how well a model encodes the regularities in color descriptions.

**Akaike Information Criterion**: AIC is derived from information theory (Akaike, 1974) and balances the model's fit to the data with the complexity of the model by penalizing a larger number of parameters. The intuition is that a smaller AIC indicates a better balance of parameters and model fit.

### 5.3 Evaluation Results

Table 1 summarizes the decision-based evaluation results.[6] We see little penalty for LUX and

[5] Specifically, the histograms are of size (90,10,10), (45,5,5), and (1,1,1) across Hue, Saturation, and Value with interpolation weights of 0.322, 0.643, and 0.035 respectively. These parameters were determined by taking the training set as 5-fold validation sets.

[6] There is a caveat to these performance measures. All of the reported numbers are for the final data subset which we discuss in Section 4.1. We choose to use a subset which did not include color labels that had less than 100 occurrences. In the English-speaking and American-citizenship subset, the rare description tail accounts for 13% of the data—Roughly one third of the tail data is unique descriptions. If the tail represents real world

| | $TOP^1$ | $TOP^5$ | $TOP^{10}$ |
|---|---|---|---|
| LUX | 39.55% | 69.80% | 80.46% |
| HM | 39.40% | 71.89% | 82.53% |
| GM | 39.05% | 69.25% | 79.99% |

Table 1: Decision-based results. The percentage of correct responses of 544,764 test-set data points are shown.

| | $-LL$ | $-LLV$ | AIC | Perp |
|---|---|---|---|---|
| LUX | 1.13*$10^7$ | 2.05*$10^6$ | 4.13*$10^6$ | 13.61 |
| HM | 1.13*$10^7$ | 2.09*$10^6$ | 4.82*$10^6$ | 14.41 |
| GM | 1.34*$10^7$ | 2.08*$10^6$ | 4.17*$10^6$ | 14.14 |

Table 2: Likelihood-based evaluation results: negative log likelihood of the data, negative log likelihood of labels given points, number of parameters, Akaike Information Criterion and perplexity of labels given color values. Parameter counts for AIC are 15751 for LUX, 315669 for HM and 5803 for GM.

GM's constrained frameworks for modeling choices. However, the differences in the table, though numerically small, are significant (by Binomial test) at $p < .02$ or less. In particular, the fact that LUX wins $TOP^1$ hints that its representations enable better generalization than HM or GM. The success of HM at $TOP^5$ and $TOP^{10}$, meanwhile, suggests that some qualitative aspects of people's use of color words do escape the strong assumptions of LUX and GM—a point we return to below.

At the same time, we draw a general lesson from the overall patterns of results in Table 1. Language users must be quite uncertain about how speakers will describe colors. Speakers do not seem to choose the most likely color label in a majority of responses; their behavior shows a long tail. These results are in line with the probabilistic models of meaning and speaker choice we have developed.

Table 2 summarizes the likelihood based metrics. GM's estimates don't fit the distribution of the test data as a whole: GM is a good model of what labels speakers give but not a good model of the points that get particular labels. By contrast, LUX tops out every row in the table. HM is flexible enough in principle to mirror LUX's predictions; HM must suffer from sparse data, given its vast number of parameters. By contrast, LUX is able to capture the distributions of speaker responses in deeper and more flexible ways by using semantics as an abstraction.

Our analysis of patterns of error in LUX suggests that LUX would best improved by more faithful models of linguistic meaning, rather than more elaborate models of subjects' choices or more powerful learning methods. For one thing, neither LUX nor the simple prototype model captures ambiguity, which sometimes arises in Munroe's data. An example is the color label *melon*, which has a multimodal distribution in the reddish-orange and green areas of color space shown in Figure 5—most likely corresponding to people thinking about the distinct colors of the flesh of watermelon, cantaloupe and honeydew. Interestingly, our model captures the more common usage.

A different modeling challenge is illustrated by the behavior of *greenish* in Figure 6. *Greenish* seems to be an exception to the general assumption that color terms label convex categories. Actually, *greenish* seems to fit the boundary of *green*—the areas that are not definitely green but not definitely not green. (Linguists often appeal to such concepts in the literature on vagueness.) This is not a convex area so, not surprisingly, our model finds a poor match. Additional research is needed to understand when it's appropriate to give meanings more complex representations and how they can be learned.

## 6 Discussion and Conclusion

Natural language color descriptions provide an expressive, precise but open-ended vocabulary to characterize real-world objects. This paper documents

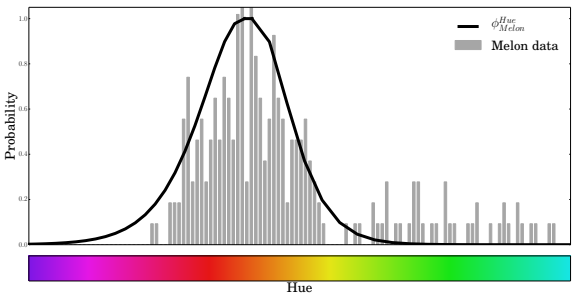


Figure 5: For the Hue dimension, the data for "melon" is plotted against the LUX model's $\phi$ curve.

circumstances, our model is only applicable 87% of the time, and thus the performance metrics should be scaled down. We do not explicitly report the scaled numbers.

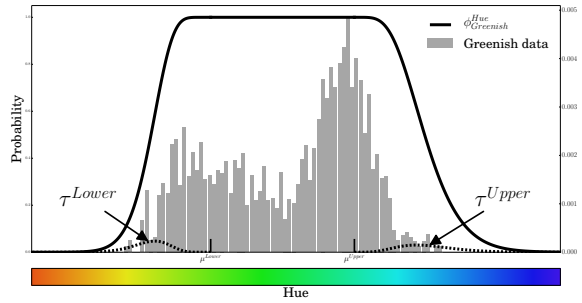


Figure 6: For the Hue dimension, the data for "greenish" is plotted against the LUX model's $\phi$ curve.

and releases Lexicon of Uncertain Color Standards (LUX), which provides semantic representations of 829 English color labels, derived from a large corpus of attested descriptions. Our evaluation shows that LUX provides a precise description of speakers' free-text labels of color patches. Our expectation therefore is that LUX will serve as a useful resource for building systems for situated language understanding and generation that need to describe colors to English-speaking users.

Our work in LUX has built closely on linguistic approaches to color meaning and psychological approaches to modeling experimental subjects. Because LUX bridges linguistic theory, psychological data, and system building, LUX also affords a unique set of resources for future research at the intersection of semantics and pragmatics of dialogue.

For example, our work explains subjects' decisions as a straightforward reflection of their communicative goals in a probabilistic setting. Our measures of availability and applicability can be seen as offering computational interpretations of the Gricean Maxims of Manner and Quality (Grice, 1975). However, these particular interpretations don't give rise to implicatures on our model—largely because our Rational Observer is so inclusive and variable in the descriptions it offers. To show this, we can analyze what an idealized hearer learns about an underlying color $x$ when the speaker uses a color term $k$: this is $P(x|k^{said})$. The model predictions are formalized in Eq. 11.

$$\begin{aligned} P(x|k^{said}) &= P(x|k^{said}, k^{true}) \\ &= \frac{P(k^{said}, k^{true}|x)P(x)}{P(k^{said}, k^{true})} \\ &= \frac{P(k^{said}|k^{true})P(k^{true}|x)P(x)}{P(k^{said}|k^{true})P(k^{true})} \\ &= \frac{\alpha_k P(k^{true}|x)P(x)}{\alpha_k P(k^{true})} = P(x|k^{true}) \end{aligned} \tag{11}$$

We apply Bayes's rule, exploiting our model assumption that the speaker says $k$ only when the speaker first judges that $k$ is true. Our model also tells us that, given that $k$ is true, the speaker's choice of whether to say $k$ depends only on the availability $\alpha_k$ of the term $k$. Simplifying, we find that the *pragmatic* posterior—what we think the speaker was looking at when she said this word—coincides with the *semantic* posterior—what we think the word is true of. Intuitively, the hearer knows that the term is true because the speaker has used the word, independent of the color $x$ the speaker is describing. Similarly, in our model of speaker choice, the speaker does not take $x$ into account in choosing one of the applicable words to say (one way the speaker could do this, for example, would be to prefer terms that were more informative about the target color $x$). Instead, the speaker simply samples from the candidates. That's why the speaker's choice reveals only what the semantics says about $x$.

Technically, this makes semantics a *Nash equilibrium*, where the information the hearer recovers from an utterance is exactly the information the speaker intends to express—in keeping with a longstanding tradition in the philosophy of language (Lewis, 1969; Cumming, 2013). By contrast, researchers such as Smith et al. (2013) adopt broadly similar formal assumptions but predict asymmetries where sophisticated listeners can second-guess naive speakers' choices and recover "extra" information that the speaker has revealed incidentally and unintentionally. The difference between this approach and ours eventually leads to a difference in the priors over utterances, but it's best explained through the different utilities that motivate speakers' different choices in the first place. Smith et al. (2013) assume speakers want to be informative; we

assume they want to fit in. The empirical success of our approach on Munroe's data motivates a larger project to elicit data that can explicitly probe subjects' communicative goals in relation to semantic coordination.

Meanwhile, our work formalizes probabilistic theories of vagueness with new scale and precision. These naturally suggest that we test predictions about the dynamics of conversation drawn from the semantic literature on vagueness. For example, in hearing a description for an object, we come to know more about the standards governing the applicability of the description. This is outlined by Barker (2002) as having a meta-semantic effect on the common ground among interlocutors. For example, hearing a yellow-green object called *yellowish green* should make objects in the same color range more likely to be referred to as *yellowish green*. We could use LUX straightforwardly to represent such *conceptual pacts* (Brennan and Clark, 1996) via a posterior over threshold parameters. It's natural to look for empirical evidence to assess the effectiveness of such representations of dependent context.

A particularly important case involves descriptive material that distinguishes a target referent from salient alternatives, as in the understanding or generation of referring expressions (Krahmer and van Deemter, 2012). Following Kyburg and Morreau (2000), we could represent this using LUX via a posterior over the threshold parameters that fit the target but exclude its alternatives. Again, our model associates such goals with quantitative measures that future research can explore empirically. Meo et al. (2014) present an initial exploration of this idea.

These open questions complement the key advantage that makes uncertainty about meaning crucial to the success of the model and experiments we have reported here. Many kinds of language use seem to be highly variable, and approaches to grounded semantics need ways to make room for this variability both in the semantic representations they learn and the algorithms that induce these representations from language data. We have argued that uncertainty about meaning is a powerful new tool to do this. We look forward to future work addressing uncertainty in grounded meanings in a wide range of continuous domains—generalizing from color to quantity, scales, space and time—and pursuing a wide range of reasoning efforts, to corroborate our results and to leverage them in grounded language use.

## Acknowledgments

This work was supported in part by NSF DGE-0549115. This work has benefited from discussion and feedback from the reviewers of TACL, Maneesh Agrawala, David DeVault, Jason Eisner, Tarek El-Gaaly, Katrin Erk, Vicky Froyen, Joshua Gang, Pernille Hemmer, Alex Lascarides, and Tim Meo.

## References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

John R. Anderson. 1991. The adaptive nature of human categorization. *Psychological Review*, 98(3):409.

Jacob Andreas and Dan Klein. 2014. Grounding language with points and paths in continuous spaces. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 58–67, June.

Chris Barker. 2002. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36.

Brent Berlin. 1991. *Basic Color Terms: Their Universality and Evolution*. Univ of California Press.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(6):1482–1493.

Stephen P. Brooks and Andrew Gelman. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 136–145.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition.

In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 128–135.

Siddhartha Chib and Edward Greenberg. 1995. Understanding the Metropolis–Hastings algorithm. *The American Statistician*, 49(4):327–335.

Jason Chuang, Maureen Stone, and Pat Hanrahan. 2008. A probabilistic model of the categorical association between colors. In *Color Imaging Conference*, pages 6–11.

Sam Cumming. 2013. Coordination and content. *Philosophers' Imprint*, 13(4):1–16.

Colin R. Dawson, Jeremy Wright, Antons Rebguns, Marco Valenzuela Escárcega, Daniel Fried, and Paul R. Cohen. 2013. A generative probabilistic framework for learning spatial language. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–8. IEEE.

David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence*, pages 747–754.

Mark D. Fairchild. 2013. *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley.

Delia Graff Fara. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, 28(1):45–81.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, June.

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Peter Gärdenfors. 2000. *Conceptual Spaces*. MIT Press.

Andrew Gelman, Gareth O. Roberts, and Walter R. Gilks. 1996. Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. Smith, editors, *Bayesian Statistics 5*, pages 599–607. Oxford University Press.

Herbert P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346.

Jeffrey Heer and Maureen Stone. 2012. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1007–1016.

John F. Hughes, Andries van Dam, Morgan McGuire, David F. Sklar, James D. Foley, Steven K. Feiner, and Kurt Akeley. 2013. *Computer Graphics: Principles and Practice (3rd Edition)*. Addison-Wesley Professional.

Gerhard Jäger. 2010. Natural color categories are convex sets. In Maria Aloni, Harald Bastiaanse, Tikitu de Jager, and Katrin Schulz, editors, *Logic, Language and Meaning - 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, volume 6042 of *Lecture Notes in Computer Science*, pages 11–20. Springer.

Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. 1977. Perplexity–a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62:S63.

Paul Kay, Brent Berlin, Luisa Maffi, William R. Merrifield, and Richard Cook. 2009. *The World Color Survey*. CSLI.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1(2):193–206.

Alice Kyburg and Michael Morreau. 2000. Fitting words: Vague words in context. *Linguistics and Philosophy*, 23(6):577–597.

Johan Maurice Gisele Lammens. 1994. *A computational model of color perception and color naming*. Ph.D. thesis, SUNY Buffalo.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*. Advance online publication. doi: 10.1093/logcom/ext059.

Daniel Lassiter. 2009. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, editors, *Vagueness in Communication - International Workshop, ViC 2009, held as part of ESSLLI 2009, Bordeaux, France, July 20-24, 2009. Revised Selected Papers*, volume 6517 of *Lecture Notes in Computer Science*, pages 127–150. Springer.

David K. Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1671–1678.

Nikolaos Mavridis and Deb Roy. 2006. Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 4690–4697. IEEE.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *SEMDIAL 2014: THE 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 107–115.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Randall Munroe. 2010. Color survey results. Online at http://blog.xkcd.com/2010/05/03/color-survey-results/.

Kimele Persaud and Pernille Hemmer. 2014. The influence of knowledge and expectations for color on episodic memory. In P Bello, M Guarini, M McShane, and B Scassellati, editors, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1162–1167.

Terry Regier, Paul Kay, and Richard S. Cook. 2005. Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, 102:8386–8391.

Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104:1436–1441.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 572–582.

Nathaniel J. Smith, Noah D. Goodman, and Michael C. Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in Neural Information Processing Systems 26*, pages 3039–3047.

Stefanie Tellex, Thomas Kollar, and Steven Dickerson. 2011a. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011b. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 1507–1514.

Terry Winograd. 1970. *Procedures as a representation for data in a computer program for understanding natural language*. Ph.D. thesis, MIT.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1050–1055.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 658–666.