Generating and Resolving Vague Color References

Timothy Meo[1] and **Brian McMahan**[2] and **Matthew Stone**[2,3]
[1]Linguistics, [2]Computer Science, [3]Center for Cognitive Science
Rutgers University
[1]New Brunswick, NJ 08901-1184, [2]Piscataway, NJ 08854-8019
firstname.lastname@rutgers.edu

Abstract

We describe a method for distinguishing colors in context using English color terms. Our approach uses linguistic theories of vagueness to build a cognitive model via Bayesian rational analysis. In particular, we formalize the likelihood that a speaker would use a color term to describe one color but not another as a function of the background frequency of the color term, along with the likelihood of selecting standards in context that fit one color and not the other. Our approach exhibits the qualitative flexibility of human color judgments and reaches ceiling performance on a small evaluation corpus.

1 Introduction

A range of research across cognitive science, summarized in Section 2, suggests that people negotiate meanings interactively to draw useful distinctions in context. This ability depends on using words creatively, interpreting them flexibly, and tracking the effects of utterances on the evolving context of the conversation. We adopt a computational approach to these fundamental skills. Our goal is to quantify them, scale them up, and evaluate their possible contribution to coordination of meaning in practical dialogue systems.

Our work extends three traditions in computational linguistics. Our approach to semantic representation builds on previous research that emphasizes the context dependence and interactive dynamics of meaning (Barker, 2002; Larsson, 2013; Ludlow, 2014). Our approach to pragmatic reasoning builds on work on referring expressions and its characterization of the problem solving involved in using vague language to identify entities uniquely in context (Kyburg and Morreau, 2000; van Deemter, 2006). Finally, we take a perceptually grounded approach to meaning, which allows us to use empirical methods to induce semantic representations on a wide scale from multimodal corpus data (Roy and Reiter, 2005; Steels and Belpaeme, 2005; McMahan and Stone, 2014).

We present our ideas through a case study of the color vocabulary of English. In particular, we study the problem solving involved in using color descriptors creatively to distinguish one color swatch from another, similar color. In our model, these descriptions inevitably refine the interpretation of language in context. We assume that speakers make choices to fulfill their communicative goals while reproducing common patterns of description. Using corpus data, we are able to quantify how representative of typical English speakers' behavior a particular context-dependent semantic interpretation is.

Our model naturally exhibits many of the preferences of previous work on vague descriptions. For example, the system avoids placing thresholds in small gaps (van Deemter, 2006), that is, in regions of conceptual space that account for little of the probability mass of possible interpretations. In such circumstances, the system prefers more specific vocabulary, where interlocutors are more likely to draw fine distinctions (Baumgaertner et al., 2012). Our approach realizes these effects by simple and uniform decision making that extends to multidimensional spaces and arbitrary collections of vocabulary.

We begin the paper by describing the semantic representation of vagueness in dialogue. Vagueness, we assume, is uncertainty about where to set the threshold in context for the concept evoked by a term. Speakers have the option to triangulate more precise thresholds by interactive strategies such as accommodation, and this helps explain how vague descriptions can be used to refer to objects precisely (van Deemter, 2006).

In Section 3, we describe our model of speakers' decisions in conversation. We focus on speak-

Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue, September 1-3, 2014, Edinburgh, U.K.

ers that aim to distinguish one thing from another; in these cases, we assume speakers aim to choose a term that's interpreted so that it fits the target and excludes the distractor, while matching broader patterns of language use.

We show how to combine the ideas in Section 4. We formalize the likelihood that a speaker would use a color term to describe one color but not another as a function of the likelihood of selecting standards to justify its application in this context, along with the background frequency of the color term. We describe an implementation of the formalism and report its the qualitative and quantitative behavior in Section 5. It works with a generic lexicon of more than 800 color terms and reaches ceiling performance in interpreting user color descriptions in the data set of Baumgaertner et al. (2012). While substantial additional research is required to explore the dynamics of vagueness in conversation, our results already suggest new ways to apply generic models of the use of vague language in support of sophisticated, open-ended construction of meaning in situated dialogue.

2 The Linguistics of Vagueness

Figure 1 shows an image from a public data set developed to study how people label images with captions (Young et al., 2014). One user chose to distinguish the dogs by calling one brown and the other tan. Another distinguished the dogs by calling one tan and the other white. Each used *the tan dog* to refer to a different dog—yet the way each described the other dog left no doubt about the correct interpretation. This variability and context dependence is characteristic of vagueness in language. The dogs in Figure 1 are borderline cases; there's no clear answer about whether they are tan or not, and speakers are free to talk of either, both, or neither of them as tan, depending on their purposes in the conversation.

In this paper, we explore the descriptive variability seen in Figure 1. How is it that speakers can settle borderline cases in useful ways to move a dialogue forward, and how can hearers recognize those decisions? We won't consider the interactive strategies that interlocutors can use to confirm, negotiate or contest potentially problematic descriptions, although that's obviously crucial for successful reference (Clark and Wilkes-Gibbs, 1986), for coordinated meaning (Steels and Belpaeme, 2005), and perhaps even for meaning itself (Ludlow, 2014). And we won't consider the way multiple descriptions constrain one another, as in Figure 1, although we expect to explain it as a side-effect of holistic interpretive processing (Stone and Webber, 1998). We see our work as a prerequisite for the model building and data collection required to address such issues.

Figure 1: A brown dog and a tan one—or a tan dog and a white one (Young et al., 2014).

In our view, the users of Young et al. (2014) are using *tan* to name color categories. Colors are visual sensations that vary continuously across a space of possibilities. Color categories are classifiers that group regions in color space together (Gärdenfors, 2000; Larsson, 2013). Color terms in English also have another sense, not at issue in this paper, where they refer to an underlying property that correlates with color, as in *red pen* (writes in red ink) (Kennedy and McNally, 2010).

Empirically, color categories seem to be convex regions (Gärdenfors, 2000; Jäger, 2010)—in fact, we model them as rectangular box-shaped regions in hue–saturation–value (HSV) space. Thus, color categories involve boundaries, thresholds or standards that delimit the regions in color space where they apply; context sensitivity can be modeled as variability in the location of these boundaries (Kennedy, 2007). For example, when we categorize the lighter dog of Figure 1 as being distinctive in its color, we must have a color category that fits this dog but not the darker one. This category will group together colors with a suitable interval of yellow hues, suitable low levels of saturation, and suitably high values on the white–black continuum. We can think of this category as one possible interpretation for the word *tan*. By contrast, categorizing the darker dog of Figure 1 as distinctively *tan* involves choosing a category with dif-

ferent thresholds for hue, saturation and value—thresholds that fit the color of the darker dog but exclude that of the lighter one.

When interlocutors use vague terms in conversation, they constrain the way others can use those terms in the future (Lewis, 1979; Kyburg and Morreau, 2000; Barker, 2002). For example, if we hear one or the other dog of Figure 1 described as *tan*, it constrains how we will interpret subsequent uses of the word *tan*. Concretely, we might update the perceptual classifier we associate with *tan* in this context so that it fits the target dog and excludes its alternative (Larsson, 2013). We see this as a case of accommodation, in the sense of (Lewis, 1979).

As speakers, we often count on our interlocutors to accommodate us (Thomason, 1990). We can use vague terms confidently as long as the distinction we aim to draw with them is clear in context and as long as our choice is sufficiently in line with the normal variation in the use of the word, and therefore uncontroversial (Thomason, 1990; van Deemter, 2006). Such criteria seem to support the speaker's choice in Figure 1 to describe either dog as *tan*—provided the speaker provides a complementary description of the other dog. At the same time, if we use language in very unusual ways, we can expect that our interlocutor may have difficulty understanding and may be reluctant to accommodate us. In other words, to use vague language effectively, speakers must be sensitive to whether their utterances update the dialogue context in a natural way.

A common idea in linguistics and philosophy is that knowledge of language associates terms with a probability distribution over categories. This distribution characterizes speakers' information about the likelihood of different possible interpretations for the term that could make sense in context (Williamson, 1996; Barker, 2002; Lassiter, 2009). In other words, vagueness amounts to uncertainty about where to draw boundaries to settle borderline cases.

Thus, when we need to settle borderline cases to generate or understand utterances like *the tan dog*, knowledge of meaning lets us quantify how likely the different resolutions are. In Figure 1, for example, knowledge of language says that *tan* can be interpreted, with a suitable probability, through categories that pick out just the lighter dog, but that *tan* can also be interpreted, with a suitable probability, through categories that pick out just the darker dog. The next section explains how to formalize the reasoning involved in assessing these probabilities, reviews one instantiation of this reasoning for learning semantics, and develops another instantiation for distinguishing colors in context.

3 Rational Analysis of Descriptions

Speakers can use language for a variety of purposes. Their decisions of what to say thus depend on knowledge of language, their communicative situation, and their communicative goals. Following Anderson (1991), rational analysis invites us to explain an agent's action as a good way to advance the agent's goals given the agent's information. When applied to communication, this approach allows us not only to derive utterances for systems but also to infer linguistic representations from utterances when we know the agent's communicative situation and communicative goals.

We apply this methodology to color descriptions in McMahan and Stone (2014). We infer linguistic representations from Randall Munroe's color corpus[1] by assuming that subjects' goals were to say true things and match a target distribution of utterances. These results are available as our Lexicon of Uncertain Color Standards (LUX). We describe this experiment in Section 3.1. We continue in Section 3.2 by creating a new model of the task of creating a distinguishing description. Here the goal is to describe one color, exclude another, and match a target distribution.

3.1 Lexicon of Uncertain Color Standards

Munroe's corpus was gathered by presenting subjects with a color patch and allowing them to freely describe it. It's not interactive language use, but we use it just to model knowledge of meaning. Like all crowdsourced data, Munroe's methodology sacrifices control over presentation of stimuli and curation of subjects' responses for sheer scale of data collection. We work with a subset of data involving 829 color terms elicited over 2.18M trials. Each description is paired with the multiset of color values on which subjects used it. We model the data in HSV space, because color categories generally differ in the *Hue* dimension.

LUX links color descriptions with context-sensitive regions in HSV color space. An example is shown in Figure 2 for the *Hue* dimension.

[1] blog.xkcd.com/2010/05/03/color-survey-results/

The plot shows a scaled histogram of subjects' responses. There is a region on the *Hue* dimension which subjects frequently described as *yellowish green* with borderline cases on either side of it.

To capture the patterns of human responses, the rational analysis approach directly models the uncertainty described in Section 2. For each color term, speakers have possible standards which can be used to partition color space; they are unsure which are at work at any point. For example, the term *yellowish green* only fits those *Hue* values which are above a minimum threshold, $\tau^{Lower,H}$ (or $\tau^{L,H}$ for short), and below a maximum threshold, $\tau^{Upper,H}$ (or $\tau^{U,H}$ for short). We estimate the distribution of possible thresholds; they are shown as the solid black lines in Figure 2.

In choosing to use the color description to fit a point x in HSV space, speakers make a semantic judgment which constrains the possible standards. The naturalness of this judgment is measured in part by the probability mass of possible standards which allow the description to be used. For example, the applicability of *yellowish green* is the probability of the color value x being between the minimum and maximum thresholds in each dimension. For a color description k, this is mathematically defined fully in Equation 1 and more compactly in Equation 2.

$$P(\tau_k^{Lower,H} < x^H < \tau_k^{Upper,H}) \times$$
$$P(\tau_k^{Lower,S} < x^S < \tau_k^{Upper,S}) \times$$
$$P(\tau_k^{Lower,V} < x^V < \tau_k^{Upper,V}) \quad (1)$$
$$= \prod_{d \in \{H,S,V\}} P(\tau_k^{L,d} < x^d < \tau_k^{U,d}) \quad (2)$$

The other factor in subjects' choices is the saliency of the color term. The saliency of color

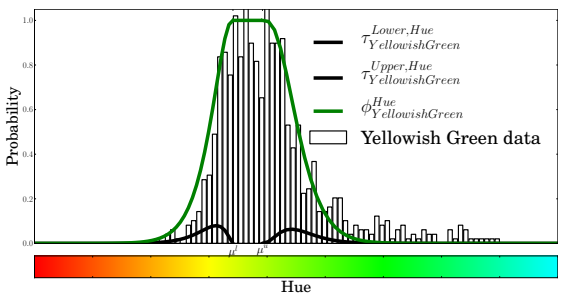

Figure 2: The LUX model for "yellowish green" on the *Hue* axis plotted against a scaled histogram of responses. The ϕ curve, the likelihood of a color counting as "yellowish green", is derived from the τ curves representing possible boundaries.

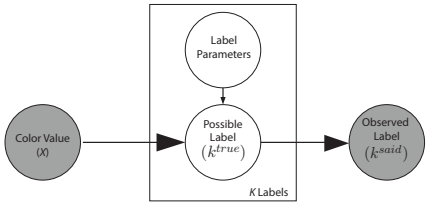

Figure 3: A Bayes Rational Observer sees a color patch. The subjective likelihood $P(k^{true}(c)|c = x)$ describes the likelihood that descriptor k is true of the current color c given that it is located *HSV* point x. The descriptor k is actually said proportional to this subjective likelihood and a weight representing how often a label is said when it is true: $P(k^{said}|k^{true}(c))$. In Munroe's data, the shaded nodes are observed.

description k, also called *availability* and written as $\alpha(k)$, is a background measure of how often the term is used when it is true. Thus, to pick a term that fits a color swatch and use language in a natural way, subjects can pick a color term according to the product of availability and subjective likelihood. Figure 3 summarizes this process in a graphical model.

In Equation 3 , we introduce a simpler notation for Equation 2 that we build on in what follows. We abbreviate $P(\tau_k^{L,d} < x^d < \tau_k^{U,d})$ as $\phi_k^d(x^d)$ and show how $\phi_k^d(x^d)$ can be defined by cases as a function of how x^d is situated with respect to the lower limit $\mu^{L,d}$ and upper limit $\mu^{U,d}$ of the threshold distributions:

$$\phi_k^d(x^d) = \begin{cases} P(x^d > \tau_k^{L,d}), & x^d \leq \mu_k^{L,d} \\ P(x^d < \tau_k^{U,d}), & x^d \geq \mu_k^{U,d} \\ 1, & otherwise \end{cases} \quad (3)$$

LUX was learned from Munroe's data by fitting the parameters of the ϕ function for each description on each dimension independently to the frequency histogram. For example, the parameters for the ϕ function for *yellowish green* in Figure 2 were fit by maximizing the probability that the bins in the data histogram were sampled from the ϕ curve with standard Gaussian noise.

3.2 Distinguishing Descriptions

Munroe's elicitation task is simple; in other settings, people have more complex communicative goals, such as unique reference. These goals modulate the link between internal semantic representation and observed speaker choice. In Munroe's

task, we assume, the speaker sampled from possible descriptive terms based on terms' availability and how likely terms were to fit the target color value. We now consider how this changes when speakers aim to differentiate between two objects.

The literature offers a key insight to get us started: referential expressions are marked as such, and the scalar structure of vague meanings gives strong constraints on how vague terms can be interpreted. For example, *the fat pig* can only refer to the fatter of two pigs in the context, a calculation that is easy to add to algorithms for referring expression generation (Kyburg and Morreau, 2000; van Deemter, 2006). However, things become substantially more complicated in the case of color, because color is multidimensional and color categories can be approximated in competing ways, as with *tan* in Figure 1.

We approach the problem probabilistically. To generate likely unique references, the speaker must sample from possible descriptive terms proportional to terms' availability, how likely terms are to fit the target, and how likely terms are to exclude a distractor. This involves integrating over all possible thresholds, to measure the probability that a description should be interpreted to include one color and exclude another. In the ordinary case where two colors are far enough apart, most thresholds work, and the approach defaults to the kinds of natural descriptions seen in descriptions of colors on their own. However, when the colors become increasingly close, general color descriptions (such as *green*) no longer are likely to signal the distinction we need, while more specific color descriptions are (such as *lime green* and *pale green*). This qualitative behavior is an important part of vague language, as observed by Baumgaertner et al. (2012). (They also suggest that accurate models of color vagueness would be necessary for good performance in difficult cases.)

The same model can inform the resolution of vague descriptions as well as generation. Resolving reference requires reasoning about how well each description applies to each of the candidate referents. We explore this reasoning for generation and understanding in the next section.

4 Algorithm and Implementation

The heart of our method is a measure of the confidence with which we can use a color term to describe a color Y and to exclude a second color Z. We will call this number the Y–but–not–Z confidence rating. This is the probability that the thresholds in context are chosen in such a way that color term k fits color Y but does not fit distractor Z. (That's $P(k^{true}(c)|c = Y) \times P(\neg k^{true}(c)|c = Z)$ in the notation of Figure 3.) To generate a term in context, we might consider each possible color label, calculate its Y–but–not–Z confidence, and finally pick a term proportional to its confidence.

We motivate our mathematical model by considering a single perceptual dimension, most easily visualized as *Hue*. In this case, the Y–but–not–Z confidence is equal to the probability that the upper and lower thresholds of that term can be set such that Y falls inside them, and Z falls outside of them. Thus each confidence rating will involve the multiplication of two values: the probabilities associated with the upper and lower boundaries.

In Figure 4, Y and Z are borderline *Hue* values; both are greener than the typical yellowish green. In this case, there's no constraint on the lower threshold; the lower threshold fits the description with probability 1. On the other hand, only the upper shaded region of Figure 4 supports a categorization of Y but not Z as yellowish green. This area is equal to $\phi(Y) - \phi(Z)$. This is the probability that the *Hue* boundaries for this color term will include Y and exclude Z. Symmetrical reasoning applies in the mirror-image case when the colors are borderline yellow.

Another case is shown in Figure 5, in which Y and Z fall on opposite sides: Y is borderline green, while Z is borderline yellow. In these *contrasting borderline* cases, it's up to the speaker whether to count Y in and Z out or vice versa, as in Figure 1. The choices can be good or bad, however,

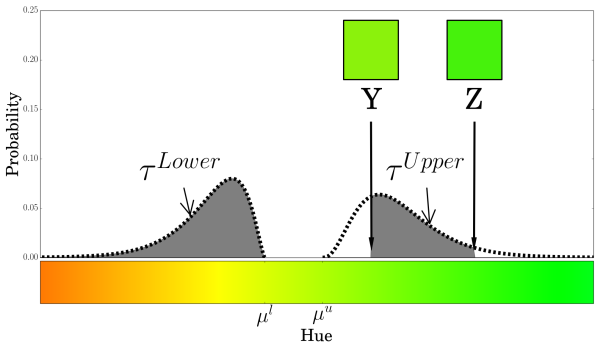

Figure 4: The thresholds that separate two nearby borderline cases cover probability $\phi(Y) - \phi(Z)$, here $0.74 - 0.05 = 0.69$.

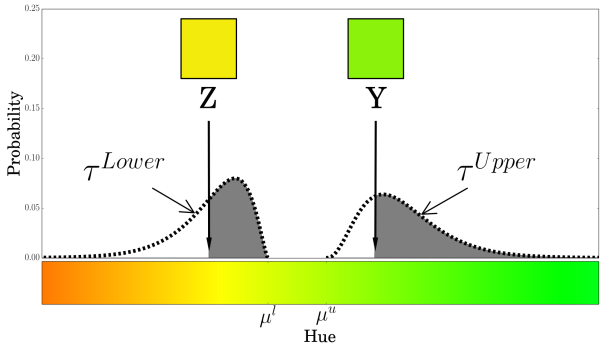

Figure 5: The thresholds that separate two contrasting borderline cases cover probability $\phi(Y) * (1-\phi(Z))$, here $0.74 * (1-0.38) = 0.46$.

because they constrain the context. The probability that the upper threshold includes Y is $\phi(Y)$. The shaded area above Z represents the probability that the lower threshold is placed such that Z is excluded; its area is equal to $1-\phi(Z)$. Thus, the Y–but–not–Z confidence rating for this case is $\phi(Y) * (1-\phi(Z))$. Again, there is a symmetrical case with the colors reversed.

Finally, if Y is not a borderline case, as in Figure 6 then Y does not constrain the thresholds at all. Thus, the Y–but–not–Z confidence rating for this case is $(1-\phi(Z))$. All three cases can be generalized to a common form, however. Let $\phi_1(Y)$ be $\phi(Y)$ if Y is a borderline case opposite Z, 1 otherwise. And let $\phi_2(Y)$ be $\phi(Y)$ if Y is a borderline case next to Z, 1 otherwise. Then all the formulas we have exhibited fit the scheme $\phi_1(Y) * (\phi_2(Y) - \phi(Z))$.

With this insight, we can extend our comparison to the three-dimensional case. The case is shown in Figure 7 for a color description k.

To calculate this probability mass we generalize Y–but–not–Z calculation to a case analysis in three

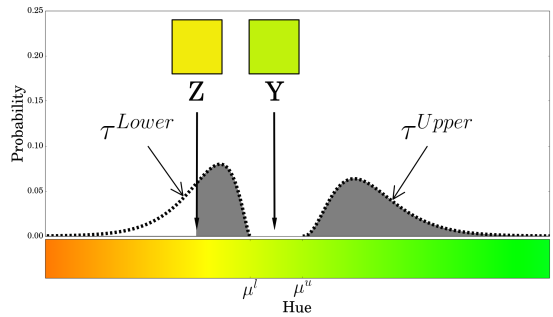

Figure 6: If Y is a clear case, we simply exclude Z, for probability $1-\phi(Z)$, here $1-0.38 = .62$.

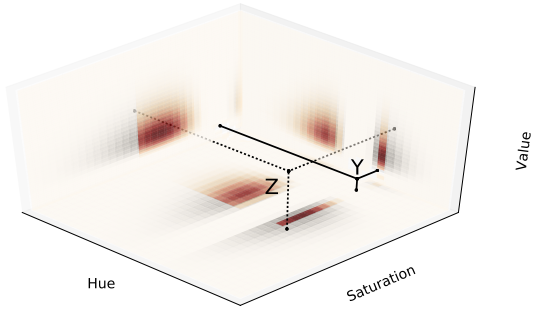

Figure 7: In the multidimensional case, solutions respect constraints from Y that are independent of Z, with probability $\phi_1(Y)$; they also select appropriate standards that affect both Y and Z, with probability $\phi_2(Y) - \phi_3(Z)$.

dimensions as shown in Equation 4.

$$\phi_1(Y) * (\phi_2(Y) - \phi_3(Z)) \tag{4}$$

In this equation, we generalize our notation to the general case as follows:

- $\phi_1(Y)$ is $\prod\phi(Y^d)$ over dimensions d where Y and Z are contrasting borderline cases
- $\phi_2(Y)$ is $\prod\phi(Y^d)$ over all other dimensions
- $\phi_3(Z)$ is $\prod\phi(Z^d)$ over all dimensions d

This expression is what we use in our implementation to calculate each color term's Y–but–not–Z confidence rating.

Given a confidence score, the evaluation is balanced by the availability of the color description.

Algorithm 1 The scoring function to compare two HSV tuples Y and Z for a single color term k

function SCORE(k,Y,Z)
  $TermA \leftarrow 1$
  $TermB \leftarrow 1$
  $TermC \leftarrow \phi_k^H(Z^H) \times \phi_k^S(Z^S) \times \phi_k^V(Z^V)$
  for each dimension d in (H,S,V) **do**
    if Y^d is on opposite side from Z^d **then**
      $TermA \leftarrow TermA \times \phi_k^d(Y^d)$
    else
      $TermB \leftarrow TermB \times \phi_k^d(Y^d)$
    end if
  end for
  $score \leftarrow TermA \times (TermB - TermC)$
  $score \leftarrow score \times \alpha(k)$
  return *score*
end function

For example, a common color term like *green* has a high availability, whereas a less frequent term, *British racing green*, has a much lower one. By weighting a term's score by its availability, we ensure that the program is less likely to generate rare color labels unless they clearly target a difficult distinction that the program needs to make.

With this score function complete, we arrive at the basic outline of our algorithm. The algorithm is shown in Algorithm 1. The *distinguish* function cycles through the dictionary, calculates the *Y*–but–not–*Z* confidence for each term *k*, and returns the results in sorted order. In the cases in which *k* describes *Z* better than it describes *Y*, the function will evaluate to a negative number. Such cases are rejected—given our model, the terms cannot describe *Y* without also describing *Z*.

5 Results

We have created an interactive visualization that allows viewers to confirm the qualitative properties of our model for themselves. Figure 8 shows a screenshot of the visualization.

Users click on either of the two color swatches on the left to select colors, which are passed to the program as two HSV triplets. The middle column then displays a list of color terms associated with those swatches; this is context-independent data pulled directly from LUX. Terms are displayed in two colors: terms that are generally good descriptions of the target color but are bad at distinguishing it from its alternative are grayed out. For example, *light green* is grayed out at the top in Figure 8, because it's such a good description of the lower swatch. The column on the right then displays the results of the generation model for the two colors. Typically, no term appears in both lists—as is true in Figure 8—because it's rare to find cases like Figure 1 where there are two plausible, competing ways to refine the meaning of a color term so as to fit one color but not the other.[2] Results are ranked by normalized confidence values; colors move up in the rankings when they more precisely distinguish the target color from its alternative. For example, *pale green* and *yellow-green* overtake the more general *spring green* as descriptions of the lower color in Figure 8.

[2]Our model does recognize a surprising difference between *lime* and *lime green* in Figure 8. This isn't a fluke: the same difference shows up in CSS color definitions for example. We suspect that *lime green* evokes the peel of the fruit but *lime* is named for the juice.

Because the colors in Figure 8 are so close, context has a strong effect in selecting differentiating descriptions. As the two colors get further apart, there's less probability mass assigned to interpretations that categorize them the same way. Under these circumstances, the differentiating color terms converge to the color terms predicted by the generic model. This recalls the heuristic of Baumgaertner et al. (2012) that basic color terms are used unless needed to distinguish. In other words, our model produces marked descriptions only when coarser terms are less reliable in distinguishing the two colors, so they are necessary to achieve the communicative goal of distinguishing the two colors. This recalls the "small gaps" constraint of van Deemter (2006).

As a first step towards quantifying the performance of the model, we got the data collected by Baumgaertner et al. (2012). They showed subjects color swatches in arrays of four, and asked subjects either to identify a particular target swatch in words (as director) or to pick the swatch that best fit a verbal description (as matcher). At issue was the ability of human matchers or various algorithms to find the original target swatch (the correct swatch) given directors' descriptions. People's success in these tasks depends on how difficult it is to distinguish the alternatives. Because problems are so variable and task dependent, there can be no universal benchmark of performance in identifying colors, but the results are helpful in understanding what we have accomplished and where further research is necessary.

Baumgaertner et al. (2012) report an analysis of 29 judgments about the interpretation of color descriptions in context across a range of difficulty levels. Their baseline algorithm, which interprets colors based on the nearest focal value in RGB space, links 23 of them to the swatch the director was instructed to describe. Of the remainder, three represent clear problems with their system. Our system, by contrast, gets all these 26 correct. The remaining three cases raise the same problems for both approaches. There seems to be one case of human error: the director is signaled to describe a brown swatch but produces *blueberry*, apparently describing the adjacent purplish-blue swatch. And two are cases of sparse data: the items *deep grey blue* and *dull salmon pink* fall out of the frequent vocabulary of Munroe's data set. The two out-of-vocabulary cases arise in the most difficult setting,

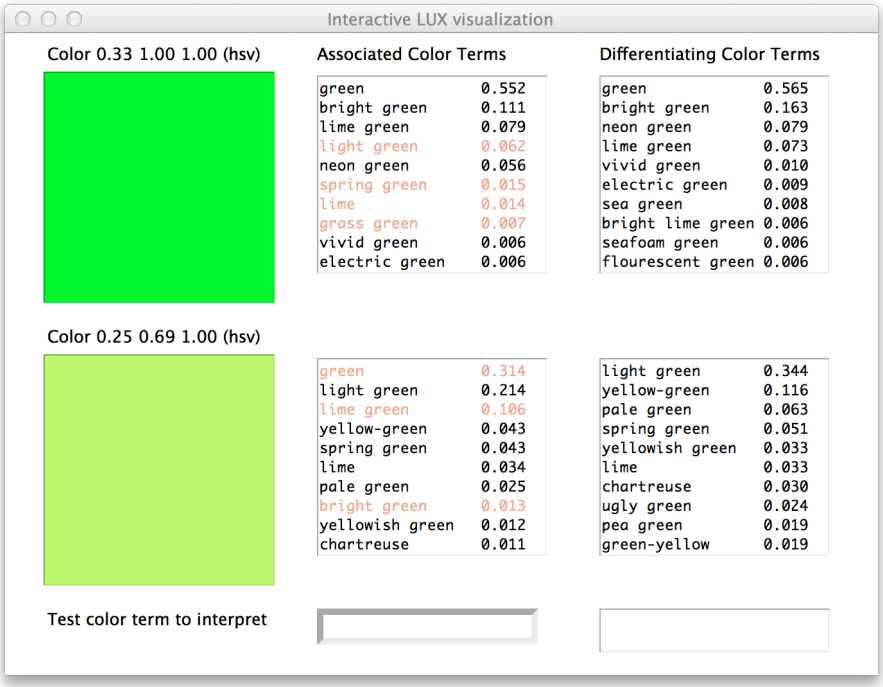

Figure 8: A screenshot of our interactive visualization, contrasting two shades of green. The system's descriptions emphasize the greater saturation and greener hue of the top color, and the lower saturation and yellower hue of the bottom color.

where directors must use low frequency terms to describe closely related colors; we get 71% right while human matchers recover the swatch signaled to the human director only 78% of the time.[3] Thus, we conclude that we need larger and more targeted data sets to distinguish the performance of our new algorithm from that of people.

Baumgaertner et al. (2012) 29 key examples are drawn from a larger elicitation experiment that produced 196 different tokens, again across a range of conditions. Our system resolves 152 correctly as written. Another 28 are out of vocabulary but closely related to terms the system would resolve correctly (differing in spelling, comparative or superlative morphology, hedges, paraphrases or other lightweight modifiers). The system gets 8 wrong as written (again, this seems to include several cases of human error); 6 are out of vocabulary and closely related to terms that the system would get wrong; and 2 are completely different from any of our vocabulary items. All the system errors are on low frequency items in situations with close distractor colors, where we've seen people also have difficulty. We were unable to find patterns of systematic error in our system.

[3]Interestingly, our system correctly resolves the alternative items *dark grey blue* and *salmon pink* in these cases. If we can deal with the productivity of low frequency descriptions, we see no obstacle to matching or even exceeding human performance.

6 Conclusion

We have explored a problem solving approach to the use of vague language. We have presented the theoretical rationale for our approach, described a broad-scale implementation, and offered a preliminary empirical evaluation.

Our work is pervasively informed by previous work on the semantics and pragmatics of dialogue. But we have not deployed or evaluated our work with interactive language use. That's an obvious and important next step.

We're excited by opportunities our work brings to assess the role of linguistic knowledge and rational problem solving in conversation. If successful, these efforts will lead to better interactive systems. But even if not, we think they will help to characterize speakers' interactive strategies, and thus to pinpoint the distinctive mechanisms that support meaning making in dialogue.

Acknowledgments

This work was supported in part by NSF DGE-0549115. Thanks to the SEMDIAL reviewers for helpful comments.

References

[Anderson1991] John R. Anderson. 1991. The adaptive nature of human categorization. *Psychological Review*, 98(3):409.

[Barker2002] Chris Barker. 2002. The dynamics of vagueness. *Linguistics and Philosophy*, 25(1):1–36.

[Baumgaertner et al.2012] Bert Baumgaertner, Raquel Fernández, and Matthew Stone. 2012. Towards a flexible semantics: colour terms in collaborative reference tasks. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 80–84. Association for Computational Linguistics.

[Clark and Wilkes-Gibbs1986] H. H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

[Gärdenfors2000] Peter Gärdenfors. 2000. *Conceptual Spaces*. MIT.

[Jäger2010] 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.

[Kennedy and McNally2010] Chris Kennedy and Louise McNally. 2010. Color, context and compositionality. *Synthese*, 174(1):79–98.

[Kennedy2007] Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

[Kyburg and Morreau2000] Alice Kyburg and Michael Morreau. 2000. Fitting words: Vague words in context. *Linguistics and Philosophy*, 23(6):577–597.

[Larsson2013] Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*.

[Lassiter2009] 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.

[Lewis1979] David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(3):339–359.

[Ludlow2014] Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press, Oxford.

[McMahan and Stone2014] Brian McMahan and Matthew Stone. 2014. A Bayesian approach to grounded color semantics. Manuscript, Rutgers University.

[Roy and Reiter2005] Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artif. Intell.*, 167(1-2):1–12.

[Steels and Belpaeme2005] Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–529.

[Stone and Webber1998] Matthew Stone and Bonnie Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of International Natural Language Generation Workshop*, pages 178–187.

[Thomason1990] Richmond H. Thomason. 1990. Accommodation, meaning and implicature. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 325–363. MIT Press, Cambridge, MA.

[van Deemter2006] Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

[Williamson1996] Timothy Williamson. 1996. *Vagueness*. Routledge, London.

[Young et al.2014] 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.