Towards a Flexible Semantics:
Colour Terms in Collaborative Reference Tasks

Bert Baumgaertner
University of California, Davis
bbaum@ucdavis.edu

Raquel Fernández
University of Amsterdam
raquel.fernandez@uva.nl

Matthew Stone
Rutgers University
matthew.stone@rutgers.edu

Abstract

We report ongoing work on the development of agents that can implicitly coordinate with their partners in referential tasks, taking as a case study colour terms. We describe algorithms for generation and resolution of colour descriptions and report results of experiments on how humans use colour terms for reference in production and comprehension.

1 Introduction

Speakers do not always share identical semantic representations nor identical lexicons. For instance, a subject may refer to a shape as a diamond while another subject may call that same shape a square (which just happens to be tilted sidewise); or someone may refer to a particular colour with *'light pink'* while a different speaker may refer to it as *'salmon'*. Regardless of these differences, which seem common place, speakers in dialogue are able to communicate successfully most of the time. Successful communication exploits interlocutors' abilities to negotiate referring expressions interactively through grounding (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989), but in many cases interlocutors can already make a good guess at their partners' intentions by relaxing the interpretation of their utterances and looking for the referent that best matches this looser interpretation. We are interested in modelling this second kind of behaviour computationally, to get a better understanding of it and to contribute to the development of dialogue systems that are able to better coordinate with their human partners.

In this paper we focus on collaborative referential tasks (akin to the classic matching tasks introduced by Krauss and Weinheimer (1966) and Clark and Wilkes-Gibbs (1986)) and take as a case study colour terms. Our focus here is not on the explicit joint negotiation of effective terms, but rather on the deployment of flexible semantic representations that can adapt to the constraints imposed by the context and to the dialogue partner's language use.

We start by describing our algorithms for generation and resolution of colour descriptions in the next section. In sections 3 and 4, we present results of experiments that investigate how humans use colour terms for reference in production and comprehension. Section 5 compares our model against the experimental data we have collected so far and discusses some directions for future work. We end with a short conclusion in section 6.

2 Reference to Colours: Our Model

Our view of how colour terms are used in referential tasks follows the basic tenets of Gricean pragmatics (Grice, 1975) and collaborative reference theories (Clark and Wilkes-Gibbs, 1986), according to which speakers and addressees tend to maximize the success of their joint task while minimizing costs.

In the domain of colour terms, we take this to mean that speakers tend use a basic colour term (e.g., *'red'* or *'blue'*) whenever this is enough to identify the target object and resort to an alternative, more specific or complex term (e.g., *'bordeaux'* or *'navy blue'*) in other contexts where the basic term is deemed insufficient. Non-basic terms can be considered more costly because they are less frequent and thus more difficult to retrieve.

Similar ideas are at the core of models for the generation of referring expressions that build on the seminal work of Dale and Reiter (1995). These approaches, however, rely on a lexicon or database where the properties of potential target objects are associated with specific, predefined terms.[1] Our aim is to develop dialogue agents that employ more flexible semantic representations, allowing them to (a) refer to target colours with different terms in different contexts, and (b) resolve the reference of colour terms produced by the dialogue partner by picking up targets that are not rigidly linked to the term in the agent's lexicon.

2.1 Algorithms

Data. To develop the generation and resolution algorithms of our agent, we used a publicly available database of RGB codes and colour terms generated from a colour naming survey created by Randall Monroe (author of the webcomic `xkcd.com`) and taken by around two hundred thousand participants.[2] This database contains a total of 954 colour terms (corresponding to the colour terms most frequently used by the participants) paired with a unique RGB code corresponding to the location in the RGB colour space which was most frequently named with the colour term in question.

We use this database as the default lexicon of our agent. Amongst the colour terms in the lexicon, we distinguish between basic and non-basic colours. We selected the following as our basic colours: red, purple, pink, magenta, brown, orange, yellow, green, teal, blue, and grey. This selection takes into account the high frequency of these terms in English and is in line with the literature on basic colour terms (Berlin and Kay, 1967; Berlin and Kay, 1991).

Resolution Algorithm. ALIN (ALgorithm for INterpretation) is given as input a scene of coloured squares and a colour term. Its output is the square it takes to be the intended target, generated as follows. Assuming the input term is in the lexicon, ALIN compares every colour in the scene to the RGB value of the input (the *anchor*). ALIN considers a colour c the intended target if, (a) c is nearest the anchor within a certain distance threshold, and (b) for any other colour c' in the scene within the given distance threshold of the anchor, c' is far enough away from both the anchor and c. We say more about distance thresholds below.

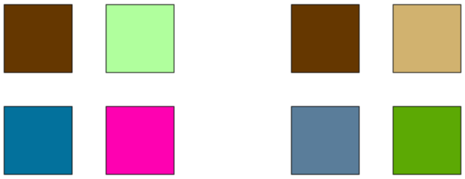

Figure 1: Two scenes with the brown square (top left in both scenes) as the target; no competitors (left scene) and one potential competitor (right scene).

Generation Algorithm. Unless there are competitors (colours relatively close to the target), GENA (GENeration Algorithm) is disposed to output a basic colour term if the target is acceptably close to a basic colour (if not, it selects the default term associated with the RGB code in the lexicon). In case there are competitor colours in the scene, if the target is a basic colour, GENA will attempt to select a non-basic colour term closest to the target but still further away from the competitor(s). If the target is not a basic colour, GENA simply selects the default term in the lexicon.

Measuring Colour Distance. We treat colours in our model as points in a conceptual space (Gärdenfors, 2000; Jäger, 2009). As a first approximation, we measure colour proximity in terms of Euclidean distances between RGB values.[3] Three variables were used to set the thresholds required by ALIN and GENA: i) *bc* is the maximum range to search for basic colours; ii) *min* is the minimum distance required between two colours to be considered minimally different; and iii) *max* is the maximum range of allowable search for alternative colours. We conducted two pilot studies to establish reasonable values for these variables, which we then set as: *bc* = 100; *min* = 25; *max* = 75.[4]

3 Experimental Methodology

We conducted two small experiments to collect data about how speakers and addressees use colour terms in referential tasks.

[1] See, however, van Deemter (2006) for an attempt to deal with vague properties such as *size* within this framework.

[2] For further details visit `http://blog.xkcd.com/2010/05/03/color-survey-results/`.

[3] We recognize Euclidean distances between RGB values assumes colour space is uniform, which is not the case in human vision (Wyszecki and Stiles, 2000). See section 5.

[4] RGB codes scaled at 0–255.

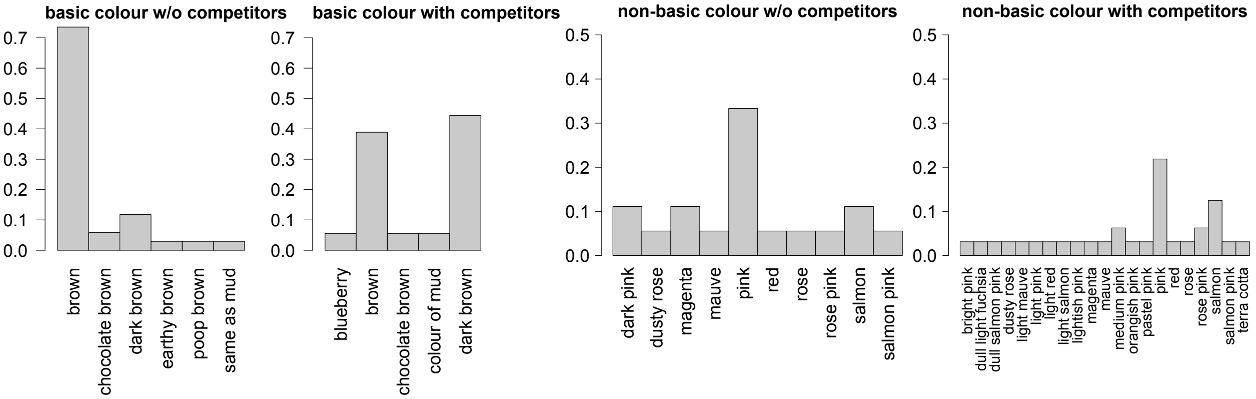

Figure 2: Sample of results from ExpA, for a basic and a non-basic colour.

Materials & Setup. We created 12 different scenes, each consisting of four solid coloured squares, one of them the target (see Figure 1 for sample scenes). Scenes were designed to take into account two parameters: basic and non-basic target colours, and without or with a competitor – a colour at a distance threshold from the target.[5] The target basic colours used were *'brown'* and *'magenta'* and the non-basic ones, *'rose'* and *'sea blue'*.[6] Each target colour appeared at least in one scene where there were no competitors.

We run a generation experiment (ExpA) and a resolution experiment (ExpB). In ExpA, participants were shown our 12 scenes and were asked to refer to the target with a colour term that would allow a potential addressee to identify it in the current context, but without reference to the other colours in the scene (to avoid comparatives such as *'the bluer square'*). In ExpB, participants were shown a scene and a colour term and were asked to pick up the intended referent. The colour terms used in this second experiment were selected from those produced in ExpA – 29 scene-term pairs in total. Each scene appeared at least twice, once with a term with high occurrence frequency in ExpA, and once or twice with one or two terms that had been produced with low frequency. To minimize chances that subjects recognize the same scene more than once, we rotated and dispersed them evenly throughout.

Participants. A total of 36 native-English participants took part in the experiments: 19 in ExpA and 17 in ExpB. Subjects for both experiments included undergraduate students, graduates students, and university faculty. Both experiments were run online.

4 Experimental Results

ExpA Generation. ExpA revealed there is high variability in the terms produced to refer to a single colour. As expected, variability of terms generated for non-basic colours was higher than for basic colours. For non-basic colours, variability of terms in scenes with competitors was higher. Figure 2 shows the different terms produced for a basic colour (*'brown'*) and a non-basic colour (*'rose'*) in scenes without and with competitors, together with the proportional frequency of each term.

For the brown square target in a scene without competitors, the basic-colour term *'brown'* was used with high frequency (72% of the time) while any other terms were used 1 or 2 times only. In scenes with competitors, *'dark brown'* had highest frequency with *'brown'* almost as much (43% vs. 40%). For the rose square target in a scene without competitors, there was also one term that stood out as the most frequent, *'pink'*, although its frequency (30%) is substantially lower to that of the basic-colour *'brown'*. In scenes with competitors there is an explosion in variation, with *'pink'* still standing out but only with a proportional frequency of 21%.

Overall, ExpA showed that speakers attempt to adapt their colour descriptions to the context and that

[5] Any colour within a Euclidean distance of 125 from the target was considered a competitor.

[6] Compositional phrases may introduce more sophisticated effects. However, the data on which our lexicon is based abstracted away from such details, treating them as simples.

there is high variability in the terms they choose to do this.

ExpB: Resolution. ExpB showed that reference resolution is almost always successful despite the variation in colour terms observed in ExpA. For the basic colours in scenes with no competitors, participants successfully identified the targets in all cases, while in scenes with competitors they did so 98% of the time. This was the case for both terms with proportionally high and low frequency.

For the non-basic colours in scenes with no competitors, the success rate in identifying the target was again 100% for both high and low frequency terms. For scenes with competitors, there were differences depending on the frequency of the terms used: for high frequency terms there were once more no resolution errors, while the resolution success rate dropped to 78% where we used terms with low proportional frequency scores. A summary of these results is shown in Table 1, together with the success rate of our resolution algorithm ALIN.

	Basic Colours				Non-basic Colours			
	high freq.		low freq.		high freq.		low freq.	
	nc	c	nc	c	nc	c	nc	c
ExpB	1	0.98	1	0.98	1	1	1	0.78
ALIN	1	0.71	1	0.71	0.5	1	0.75	0.71

Table 1: Resolution success rate by human participants and ALIN in scenes without and with competitors (nc/c).

5 Discussion

The data we collected allows us to make informative comparisons between humans and our model in collaborative reference tasks. Although we do not believe the data is sufficient for an evaluation, the comparison illuminates how the model can be refined and the setup required for a proper evaluation.

Regarding resolution, we note that an algorithm that rigidly associates colours and terms would have successfully resolved only 4 of the 29 cases, 3 of which were basic colours with no distractors – a 7.25% success rate. In our scenarios with four potential targets, a random algorithm would have an average success rate of 25%. ALIN is closer to our human data (see Table 1), though anomalies exist. One problem is the lack of compositional semantics in our current model. ALIN failed to resolve complex phrases like *'dull salmon pink'* and *'deep gray blue'*, which were terms produced by humans for non-basic colours with competitors, simply because the terms were not in the agent's lexicon. Other anomalies seem to be consequences of taking Euclidean distances over RGB values, which may be too crude. In the future, our intent is to convert RGB values to Lab values and then use Delta-E values to measure distances. First, however, we need a more sophisticated analysis of the thresholds that we used for ALIN and GENA.

As for generation, given the amount of variation observed in the terms produced by our subjects, it is not clear how human performance ought to be compared to GENA's. For instance, in scenes with competitors, GENA produced *'reddish brown'* for the basic colour *'brown'* and *'coral'* for the non-basic colour *'rose'*. These did not appear in our human-generated data but still seem to our lights reasonable descriptions. GENA also produced *'gray'* to refer to *'rose'* in a different scene, which seems less appropriate and may be due to our current way of calculating colour distances and setting up the thresholds.

We believe that instead of comparing GENA's output to human output, it makes more sense to evaluate GENA by testing how well humans can resolve terms produced by it. We intend to carry out this evaluation in the future.

6 Conclusions

We have focused on the specific case of colours where speakers differ in the referring expressions they generate, but addressees are nevertheless able to relax the interpretations of the expressions in order to coordinate. We believe this implicit adaptability is part of our semantic representation more broadly. The case of colour provides us with a starting point for studying and modelling computationally this flexibility we possess.

Acknowledgements

This work has been partially supported by grant 632.002.001 from the Netherlands Organisation for Scientic Research (NWO) and by grant IIS-1017811 from the National Science Foundation (NSF).

References

Brent Berlin and Paul Kay. 1967. *Universality and evolution of basic color terms*. Laboratory for Language-Behavior Research.

Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution*. Univ of California Pr.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Herbert H. Clark and Donna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 18:233–266.

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

Paul Grice. 1975. Logic and conversation. In D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson, Encino, California.

Gerhard Jäger. 2009. Natural color categories are convex sets. In *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, pages 11–20. Springer.

Robert Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346.

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

Günter Wyszecki and Walter S. Stiles. 2000. *Color science: concepts and methods, quantitative data and formulae*. Wiley Classics Library.