Computational Semantics and Pragmatics

Autumn 2013

Raquel Fernández
Institute for Logic, Language & Computation
University of Amsterdam

Outline

- Last week:
  - ∗ Classic approaches to GRE and some of their extensions generation of an “optimal” but human-like description
    - ▶ no more / less information than required – by the speaker?
    - ▶ no more / less information than required – by the addressee?
- Today:
  - ∗ Reference to colours (Baumgaertner et al. 2012)
  - ∗ Interactive, collaborative reference
- Tomorrow:
  - ∗ Discussion of Jordan & Walker (2005)
  - ∗ Other computational approaches to interactive referring

Referring to Colours

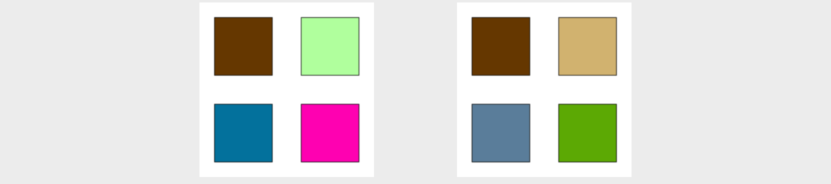
- Project carried out by Bert Baumgaertner, at the time PhD student at University of California, Davis
- project started at Rutgers with Matthew Stone
- it continued at Amsterdam in autumn 2011

Bert Baumgaertner, Raquel Fernández, and Matthew Stone (2012). Towards a Flexible Semantics: Colour Terms in Collaborative Reference Tasks. In *Proc. First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada.

Issues We Wanted to Model (I)

Speakers follow basic pragmatic principles when referring to colours: say enough but not more than is required

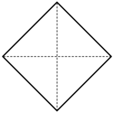
- tend to use a basic colour term whenever this is enough
- but resort to alternative terms (e.g., *'bordeaux'* or *'navy blue'*) in contexts where the basic term is deemed insufficient.
  - ∗ non-basic terms can be considered more costly because they are less frequent and thus more difficult to retrieve.



Issues We Wanted to Model (II)

Dialogue participants do not always share identical semantic representations nor identical lexicons.

But they are able to communicate successfully most of the time.

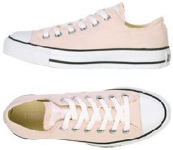


```
A: a diamond
B: ok
```

[*A must mean the tilted square*]

```
A: the salmon shoes
B: ok
```

[*A must mean those pink shoes*]



Addressees are able to relax the interpretation of the speaker's terms and look for the referent that best matches this looser interpretation.

Aims

Can we implement an artificial dialogue agent that employs flexible semantic representations, allowing it to

- refer to target colours with different terms in different contexts
- 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.

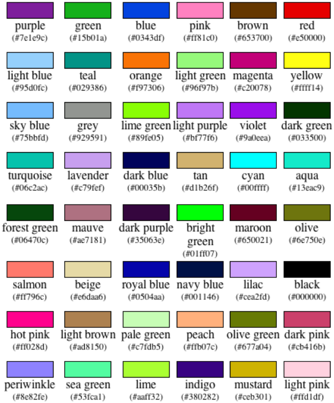
The first element we need is a model of the agent's lexicon.

Our Agent's Lexicon

Data: publicly available database of RGB codes and colour terms created by Randall Monroe (author of the webcomic `xkcd.com`)

- colour naming survey taken by around two hundred thousand participants
- 954 colour terms (the most frequently used by the participants)
- paired with a unique RGB code (location in the RGB colour space most frequently named with the colour term in question.)

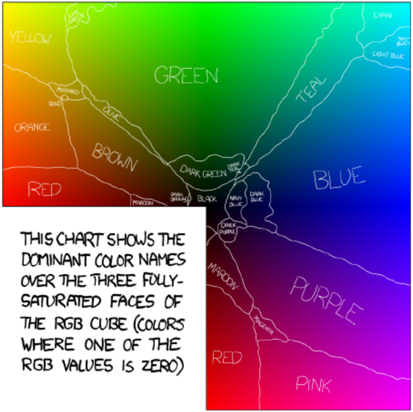
http://blog.xkcd.com/2010/05/03/color-survey-results/



purple (#7e1e9c)	green (#15b01a)	blue (#0343df)	pink (#ff81c0)	brown (#653700)	red (#e50000)
light blue (#95d0fc)	teal (#029386)	orange (#f97306)	light green (#96f97b)	magenta (#c20078)	yellow (#ffff14)
sky blue (#75bbfd)	grey (#929591)	lime green (#89fe05)	light purple (#bf77f6)	violet (#9a0eea)	dark green (#033500)
turquoise (#06c2ac)	lavender (#c79fef)	dark blue (#00035b)	tan (#d1b26f)	cyan (#00ffff)	aqua (#13eac9)
forest green (#06470c)	mauve (#ae7181)	dark purple (#35063e)	bright green (#01ff07)	maroon (#650021)	olive (#6e750e)
salmon (#ff796c)	beige (#e6daa6)	royal blue (#0504aa)	navy blue (#001146)	lilac (#cea2fd)	black (#000000)
hot pink (#ff028d)	light brown (#ad8150)	pale green (#c7fdb5)	peach (#ffb07c)	olive green (#677a04)	dark pink (#cb416b)
periwinkle (#8e82fe)	sea green (#53fca1)	lime (#aaff32)	indigo (#380282)	mustard (#ceb301)	light pink (#ffd1df)

xkcd

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



THIS CHART SHOWS THE DOMINANT COLOR NAMES OVER THE THREE FULLY-SATURATED FACES OF THE RGB CUBE (COLORS WHERE ONE OF THE RGB VALUES IS ZERO)

Colour Model

Amongst the 954 colour terms in the lexicon, we pick up 10 which we consider basic colours:

- red, purple, pink, magenta, brown, orange, yellow, green, teal, blue, and grey.
- the choice is based on their high frequency in English and is consistent with studies by Berlin and Kay (1967,1991).

Berlin & Kay (1991). Basic color terms: Their universality and evolution. UC Press.

We treat colours as points in a conceptual space

- RGB dimensions (ranging from 0 to 255)
- we measure colour proximity in terms of Euclidean distances between RGB values.

Gärdenfors (2000). Conceptual Spaces. MIT Press, Cambridge.

Generating & Resolving Colour References

We want to use the knowledge encoded in our agent's lexicon in flexible ways for resolution (interpretation) and generation (production) of colour terms in referential tasks.

Our algorithms make use of three thresholds:

- *min*: minimum distance required for two colours to be considered different.
- *max*: maximum range of allowable search for alternative colours
- *compdist*: distance range within which a colour is considered a competitor

Current tentative settings: $min = 25$; $max = 100$; $compdist = 125$

Resolution Algorithm

```
get inputTerm
get sceneColours
if inputTerm $\notin$ lexicon
  return 'I don't know this colour'
else
  anchor = term2rgb(inputTerm)
  Dist = { }
  for each c $\in$ sceneColours
     Dist = Dist $\cup$ dist(c, anchor)
  for each c $\in$ sceneColours
     if dist(c, anchor) = argmin(Dist) & dist(c, anchor) < $max$
       bestTarget = c
     else return 'I can't find anything of that colour'
  if bestTarget is defined
    enoughDiff = true
    for each (c != bestTarget) $\in$ sceneColours
       if dist(c, anchor) < $min$ & dist(c, bestTarget) < $min$
         enoughDiff = false
    if enoughDiff = true
       return bestTarget
    else return 'I can't tell'
```

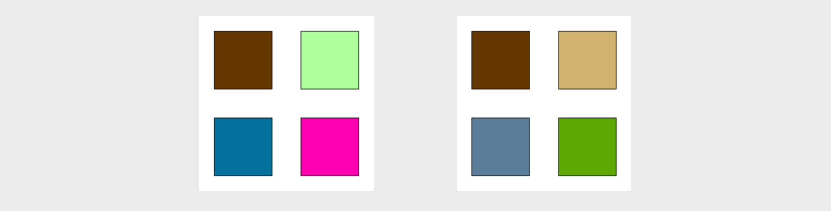
Generation Algorithm

```
get targetColour
get sceneColours
Competitors = {c ∈ sceneColours | dist(c,targetColour) < compdist}
if Competitors = { }
  b = basic colour closest to targetColour
  if dist(b, targetColour) < max
    return colour2term(b)
  else return colour2term(targetColour)
if Competitors != { }
  if targetColour ∈ basicColours
    sep = 0 & foundMoreSpecific = false
    while (findMoreSpecific = false) & (sep > max) do
      for each col ∈ lexicon
        if dist(col, targetColour) = sep
          closeColour = col
      for each comp ∈ Competitors
        if dist(comp, closeColour) > dist(comp, targetColour)
          findMoreSpecific = true
          return colour2term(closeColour)
        else sep++
  elsif targetColour ∉ basicColours
    return colour2term(targetColour)
```

What do people actually do?

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

- 12 different scenes, each with 4 solid coloured squares, one being the target.
- Scenes generated according to two parameters:
  - ∗ basic vs. non-basic target colour (*brown* or *magenta* vs. *rose* or *blue*)
  - ∗ with or without competitors: colours at a distance threshold of 125



Experiments: Generation & Resolution

Generation (ExpA):

- participants were shown a series of scenes each with a target
- they 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)

Resolution (ExpB):

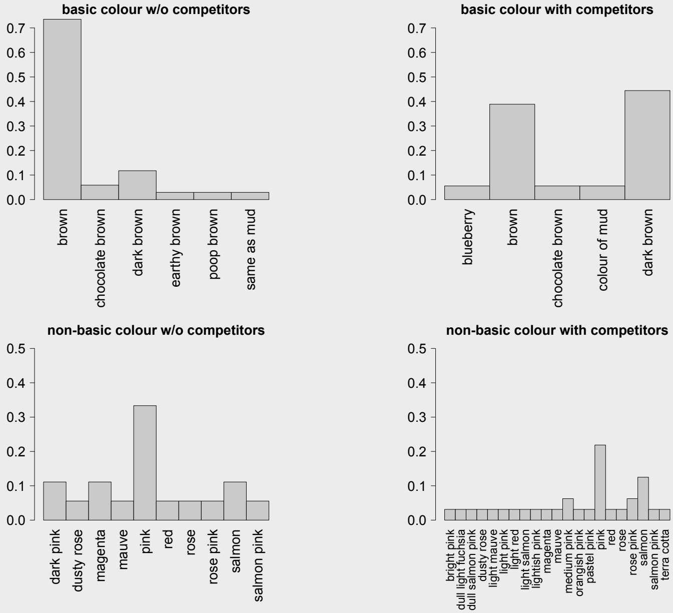
- participants were shown a series of scenes each with a colour term
- they were asked to pick up the intended referent
- the colour terms used were selected from those produced in ExpA (more details later)

The two experiments were run online, with 36 native-English participants: 19 in ExpA and 17 in ExpB.

ExpA Generation: Results

ExpA showed that speakers attempt to adapt their colour descriptions to the context and that there is high variability in the terms they choose to do this.

- higher variability of terms for non-basic than for basic colours
- for non-basic colours, higher variability of terms in scenes with competitors


basic colour w/o competitors
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
brown
chocolate brown
dark brown
earthy brown
poop brown
same as mud
basic colour with competitors
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
blueberry
brown
chocolate brown
colour of mud
dark brown
non-basic colour w/o competitors
0.0
0.1
0.2
0.3
0.4
0.5
dark pink
dusty rose
magenta
mauve
pink
red
rose
rose pink
salmon
salmon pink
non-basic colour with competitors
0.0
0.1
0.2
0.3
0.4
0.5
bright pink
dull light fuchsia
dull salmon pink
dusty rose
light mauve
light pink
light red
light salmon
lightish pink
magenta
mauve
medium pink
orangish pink
pastel pink
pink
red
rose
rose pink
salmon
salmon pink
terra cotta

ExpB Resolution

The colour terms used in ExpB were selected from those produced in ExpA

- for each scene, we selected one term produced with high frequency and one or two terms produced with low frequency
- 29 scene-term pairs in total; each scene appeared at least twice (rotated and dispersed).

ExpB Resolution: Results

Reference resolution is almost always successful despite the high variation of terms observed in generation.

- Basic colours:
  - ∗ without competitors: participants successfully identified the targets in all cases (100% success rate)
  - ∗ with competitors: 98% success rate
  - ∗ the same results for terms with proportionally high and low freq.
- Non-basic colours:
  - ∗ without competitors: 100% success in all cases (low/high freq.)
  - ∗ with competitors: differences as an effect of frequency
    - ▶ terms produced with high frequency: no resolution errors
    - ▶ low frequency terms: resolution success rate dropped to 78%

Comparing Our Model to the Human Data

- The experimental data allows us to make informative comparisons between humans and our model.
- The data is not sufficient for a proper evaluation
- but the comparison illuminates how the model can be refined and what the setup required for a proper evaluation would be.

Comparing resolution: success rate

	Basic Colours				**Non-basic Colours**			
	high freq.		low freq.		high freq.		low freq.	
%	nc	c	nc	c	nc	c	nc	c
Humans ExpB	100	98	100	98	100	100	100	78
Resolution algorithm	100	71	100	71	50	100	75	71

c = competitors
nc = no competitors

- An agent 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.
- A random algorithm would have an average success rate of 25% (four potential targets)
- Our algorithm is closer to human performance, but there are problems.

Some problems of our current model

- Lack of compositional semantics
  - ∗ some complex phrases produced by humans for non-basic colours with competitors: *'dull salmon pink'* and *'deep gray blue'*
  - ∗ resolution failed because they were not in the agent's lexicon
- Euclidean distances over RGB values
  - ∗ seem too crude
  - ∗ a better approach closer to human perception: Lab colour model with distance over Delta-e values
- Thresholds
  - ∗ we need a more systematic and empirically motivated way to set the thresholds used by the algorithms

How to evaluate 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 the automatic generation.

- in scenes with competitors, our agent produced *'reddish brown'* for the basic colour *'brown'* and *'coral'* for the non-basic colour *'rose'*, which did not appear in our human data but seem fine
- the agent also produced *'gray'* to refer to *'rose'* in a different scene, which seems less fine. . .

More appropriate evaluation of generation algorithm:

- test how well humans can resolve terms produced by it

Summing-up & possible future directions

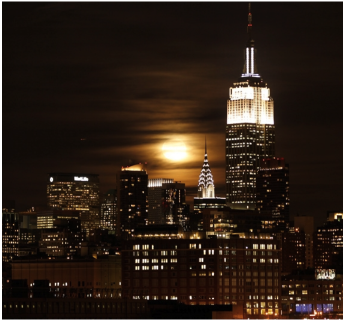
- Our aim was to model implicit processes of adaptation in resolution and generation of referring expressions, focusing on the specific case of colours
- The results of our small-scale experiments show that speakers differ greatly in the expressions they generate, but addressees are nevertheless able to coordinate.
- So far we have modelled ad-hoc adaptation to the context and to the partner, but it would be very interesting to explore how the lexicon may get modified as a result of coordination (learning).
- Finally, colours were taken as a case study, but one could try to extend the approach to other types of expressions.

Referring in Interactive Settings

So far, the referring tasks we have seen are not really interactive

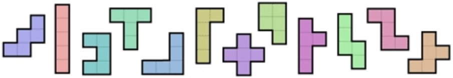
- classic GRE deals with one-shot descriptions
- even in the colours model, the focus is on individual processes of generation and resolution.

These processes are interesting and important in their own right, but referring in conversation is in fact very different . . .



A: I really like the Chrysler building.
B: The one that looks like a church?
A: You see the big Empire State buidling?
B: Yep.
A: The Chrysler is the smaller one to the left of it.
B: The one that looks like a church.
A: Yeah, I guess so.
B: It's nice, especially with the lights on.

[based on an example by Alexander Koller]



B: it’s a block of three . and then one tagged on . to the edge
A: oh it’s like . . a symmetrical L and then another two blocks . attached on to another end kind of thing
B: What? [*laughter*]
A: Okay, uhm you’ve got . . uh (t- + two) blocks
B: Yeah.
A: Uhm and then on the end of those two blocks
B: Yeah.
A: you’ve got .. . another . block (it’s like + it’s) kind of making an L
B: u:hm.
A: and then . . on that block . on that edge . uhm
B: I think I know what you’re talking about, so there’s three blocks up and one block across but in the middle block . of the one that’s going up there’s one sticking out [. . .]
A: One by one block that’s been taken out and it’s been moved
B: Yes and this has been put in the middle. Yeah yeah yeah yeah.
A: In the middle. Yeah?
B: Yeah, got it.
A: Yeah, OK.

Referring in Interactive Settings

- speakers don't get only one chance to produce a description – they can reformulate
- they receive online feedback from their addressees
- addressees themselves contribute to the referring process
- referring expressions do not emerge from solitary choices of the speaker (cf. Gricean maxims), but from an interactive process by speaker and addressee.
- speakers and addressees can agree on a description for a referent during the referring process – what works for a dyad may not work for another one

⇒ Referring is a joint process where speakers and addressees try to minimize collaborative effort.

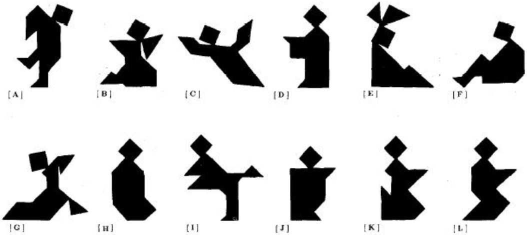
Clark & Wilkes-Gibbs (1986) Referring as a collaborative process. *Cognition*, 22:1-39.

Brennan & Clark (1996) Conceptual Pacts and Lexical Choice, *Journal of Experimental Psychology*, 22(6):1482–1493.

Matching Referring Tasks

The classic *"Tangram experiments"* by Clark & Wilkes-Gibbs:

- matching referring task: an instruction giver (director) and an instruction follower (matcher)
- the task is to get the matcher identify the tangram figures
- the task is repeated (in different orders) over several trials



This facilitates investigation of the referring process as participants accumulate common ground and precedents for referring expressions.

Referring as a Collaborative Process

Basic exchange:

(1) A: Number 4's *the guy leaning against the tree*.
B: Okay.

Refashionings:

(2) A: OK, the next one is the rabbit.
B: Uh–
A: That's asleep, you know, it looks like it's got ears and a head pointing down?
B: Okay.

(3) A: Um, the third one is the guy reading with, holding his book to the left.
B: Okay, kind of standing up?
A: Yeah.
B: Okay.

Basic exchanges occur seldom on early trials (6%) but often on later trials (84%). Refashionings decline in later trials once a RE has been mutually established.

Minimizing Collaborative Effort

- Clark & Wilkes-Gibbs' Principle of Least *Collaborative* Effort

  > *"Our proposal is that speakers and addressees try to minimize collaborative effort, i.e. the work both speakers and addressees do from the initiation of the reference process to its completion"*

- There is a trade-off in effort between initiating an expression and refashioning it: the more effort the speakers put in the initial expression, the less refashioning it is likely to need.
- Initial expressions are not always optimal due to time pressure, complexity, ignorance, ...
- Speakers deal with these constraints minimizing collaborative effort with repairs, instalments, and trial and error.
- Addressees minimize collaborative effort by indicating quickly and informatively what is needed for mutual acceptance.

Establishing Conceptual Pacts

When speakers and addressees arrieve at a successful expression (*ground* a reference), they create a *conceptual pact*, a temporary agreement about a conceptualisation for a particular entity.

A: A docksider.
B: A what?
A: Um.
B: Is that a kind of dog?
A: No, it's a kind of um leather shoe, kinda pennyloafer.
B: Okay, okay, got it.

⇒ Thereafter "the pennyloafer"

Conceptual pacts

- overwrite quantity maxims: they will continue to call it *'the pennyloafer'* even when it does not need to be distinguished from other shoes
- are partner-specific: they will do so only when interacting with the dialogue partner with whom the expression had been grounded.

To Do

- Readings for tomorrow:

Jordan & Walker (2005) Learning Content Selection Rules for Generating Object Descriptions in Dialogue, *Journal of Artificial Intelligence Research*, 24:157–194.

- Have a look at homework #2 (due Tue 1 Oct before class)