Perceptual, Conceptual, and Frequency Effects on Error Patterns in English Color Term Acquisition

Barend Beekhuizen & Suzanne Stevenson

Department of Computer Science
University of Toronto

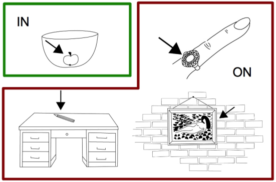
CogACLL, 18 September 2015

Overview

- What causes children to make errors in word-meaning acquisition?
- Typological Prevalence Hypothesis
- Earlier work for space
- Extension to color

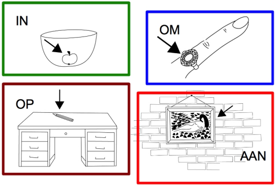
Typological Prevalence Hypothesis

- [?]:The more languages group two situations under the same linguistic label, the more cognitively natural that grouping is and hence, the easier to acquire for children



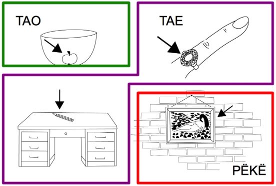
Typological Prevalence Hypothesis

- [**?**]:The more languages group two situations under the same linguistic label, the more cognitively natural that grouping is and hence, the easier to acquire for children



Typological Prevalence Hypothesis

- [**?**]:The more languages group two situations under the same linguistic label, the more cognitively natural that grouping is and hence, the easier to acquire for children



Beekhuizen, Fazly & Stevenson (2014)

- This hypothesis in a computational model
- Extract semantic space from linguistic elicitations
- Dutch children: use *op* for *aan*-situations (overextension).
- Follows from semantic space

Beekhuizen, Fazly & Stevenson (2014)

- This hypothesis in a computational model
- Extract semantic space from linguistic elicitations
- Dutch children: use *op* for *aan*-situations (overextension).
- Follows from semantic space

	English		Dutch				Tiriyo		
	on	in	op	aan	om	in	tao	tae	pëkë
apple in bowl	0	1	0	0	0	1	1	0	0
ring on finger	1	0	0	0	1	0	0	1	0
pen on table	1	0	1	0	0	0	0	1	0
painting on wall	1	0	0	1	0	0	0	0	1

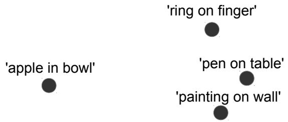
Beekhuizen, Fazly & Stevenson (2014)

- This hypothesis in a computational model
- Extract semantic space from linguistic elicitations
- Dutch children: use *op* for *aan*-situations (overextension).
- Follows from semantic space

	ring on finger	pen on table	painting on wall
apple in bowl	1	1	1
ring on finger		0.5	0.25
pen on table			0.4

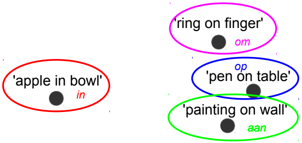
Beekhuizen, Fazly & Stevenson (2014)

- This hypothesis in a computational model
- Extract semantic space from linguistic elicitations
- Dutch children: use *op* for *aan*-situations (overextension).
- Follows from semantic space



Beekhuizen, Fazly & Stevenson (2014)

- This hypothesis in a computational model
- Extract semantic space from linguistic elicitations
- Dutch children: use *op* for *aan*-situations (overextension).
- Follows from semantic space



- Study effect of typological prevalence in other domains
- Here : Color
- Our focus: many overextensions: why?
    - Overextension = use of word to express a meaning for which adult speakers have another word (e.g., *op* for *aan*).
- Explanation in terms of typological prevalence?

[?]

- 591 English-speaking children, age 6-12.
- Shown 8 color chips each, for colors BLACK, WHITE, RED, YELLOW, GREEN, BLUE, ORANGE, and PURPLE
- Results:
  - BLACK, WHITE, RED, and BLUE: hardly any errors;
  - GREEN and YELLOW: a few early errors;
  - ORANGE: somewhat haphazard, persistent errors;
  - PURPLE: persistent errors, mostly *blue* (but not *purple* for BLUE!)

- What causes this error pattern?
  - Usual suspect #1: color term frequency
  - Usual suspect #2: perceptual features of colors [**?**]
  - New: Typological prevalence of color groupings?
- Approach:
  - Cognitive model parametrizing these possible factors
    - perceptual features $\rightarrow$ part of meaning space,
    - typological prevalence $\rightarrow$ part of meaning space,
    - frequency $\rightarrow$ part of input-item sampling procedure.
  - Give model Bateman's color chips and ask for most likely color term.
  - Evaluate fit with Bateman's observed error pattern given various parameter settings.

- Do the perceptual features of the colors play a role?
- CIELab space
- If `+perc`, every exemplar has as a coordinate in this space

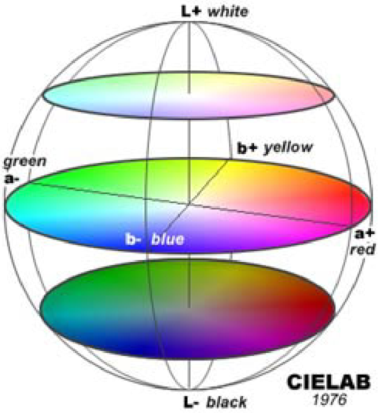
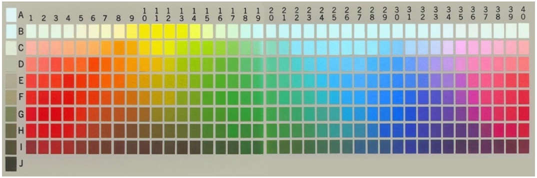


Figure: CIELab space

- Does typological prevalence play a role?
- Same approach as outlined earlier:

- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])



- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])

Count matrix for English

	white	*pink*	*orange*	...	*purple*
chip A1	15	0	0	...	0
chip A2	2	13	0	...	0
⋮					
chip I40	0	0	0	...	15

- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])
  - per language: distance matrix between all color chips,

Distance matrix for English

	chip B1	chip C1	chip D1	...	chip I40
chip A1	0.81	0.87	0.98	...	1
chip B1		0.26	0.42	...	0.96
⋮					
chip H40				...	0.81

- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])
  - per language: distance matrix between all color chips,
  - sum distance matrices for all languages,

Distance matrix for all languages

	chip B1	chip C1	chip D1	...	chip I40
chip A1	120.4	122.1	136.8	...	142.0
chip B1		73.6	82.1	...	128.1
⋮					
chip H40				...	112.6

- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])
  - per language: distance matrix between all color chips,
  - sum distance matrices for all languages,
  - apply PCA, use components with Eigenvalue > 1,

PCA coordinates for all color chips

	PCA1	PCA2	PCA3	...	PCA330
chip A1	2.4	-4.2	3.8	...	0.0
chip B1	2.7	-1.9	1.0	...	0.0
⋮					
chip H40	-4.2	2.2	3.2	...	0.0

- Does typological prevalence play a role?
- Same approach as outlined earlier:
  - get count matrix per language from linguistic elicitations (World Color Survey [**?**])
  - per language: distance matrix between all color chips,
  - sum distance matrices for all languages,
  - apply PCA, use components with Eigenvalue > 1,
- If $+\texttt{conc}$, all exemplars have a coordinate in this space.

- Does frequency in CDS play a role?

- Does frequency in CDS play a role?
- Sampling procedure: we sample iteratively pairs of a color term t and a situation s,
  - where s is the vector of the `perceptual` and/or `conceptual` coordinates.
- $P(s, t) = P(s|t)P(t)$
  - $P(s|t) = \frac{n(t,s)}{\sum_{s' \in S} n(t,s')}$

- Does frequency in CDS play a role?
- Sampling procedure: we sample iteratively pairs of a color term t and a situation s,
  - where s is the vector of the `perceptual` and/or `conceptual` coordinates.
- $P(s,t) = P(s|t)P(t)$
  - $P(s|t) = \frac{n(t,s)}{\sum_{s' \in S} n(t,s')}$
- Two conditions:
  - `relative`: $P(t)$ is relative frequency of color terms in CDS, [**?**],
  - `uniform`: $P(t)$ is uniform.

Overview

- Train learning model on iteratively sampled t, s pairs

t	s
red	$\underbrace{52.4, 0.8, 0.2,}_{\texttt{perc}} \underbrace{0.83, 0.23, 0.41, 0.03}_{\texttt{conc}}$

- Every 10 input items (test moment)
  - give model the 8 colors of [?] (using focal colors [?] represented as an s)
  - ask model for most likely color term t
- Evaluate fit between predicted responses and observed responses

Gaussian Naive Bayes (GNB)

- Centroid learner
- Learns Gaussians over the dimensions of the situations (perceptual and conceptual) from available data
- Categorizes test item on the basis of Maximum A Posteriori probability

Generalized Context Model (GCM)

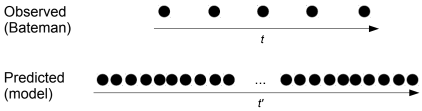
- Exemplar learner [**?**]
- Categorizes test item on the basis of similarity to all stored exemplars

- Simulation runs for 1000 input items
- I.e. 100 test moments for the 8 color chips

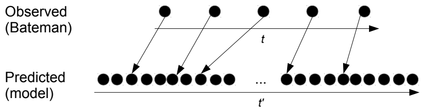
- Simulation runs for 1000 input items
- I.e. 100 test moments for the 8 color chips
- 30 simulations per combination of parameter settings:
  - `features = {perc&conc, perc, conc}` $\times$
  - `frequency = {relative, uniform}` $\times$
  - `model = {GNB, GCM}`

- Comparing test moment with observed moment:
  - 30 simulations: distribution over color terms
  - *n* children: distribution over color terms
  - take Euclidean distance between them (**error**).

- Comparing test moment with observed moment:
  - 30 simulations: distribution over color terms
  - *n* children: distribution over color terms
  - take Euclidean distance between them (**error**).
- However, 100 test moments and only 5 age bins.
- So, align predicted with observed data



- Comparing test moment with observed moment:
  - 30 simulations: distribution over color terms
  - *n* children: distribution over color terms
  - take Euclidean distance between them (**error**).
- However, 100 test moments and only 5 age bins.
- So, align predicted with observed data
- Solution: 5 test moments that have lowest distance to 5 age bins (over all 8 colors)
- Constraint: linearly ordered



parameter	setting	mean error
features ***	perc&conc	$\mu = 0.015$
	perc	$\mu = 0.020$
	conc	$\mu = 0.354$
frequency	relative	$\mu = 0.130$
	uniform	$\mu = 0.130$
model *	GCM	$\mu = 0.120$
	GNB	$\mu = 0.139$

Interpretation

- No effect of frequency: cf. [?]?
- Small effect of model
- Effect of features: perc, perc&conc $>$ conc:
  - English is natural
  - Children too old
  - Color is easier domain then space

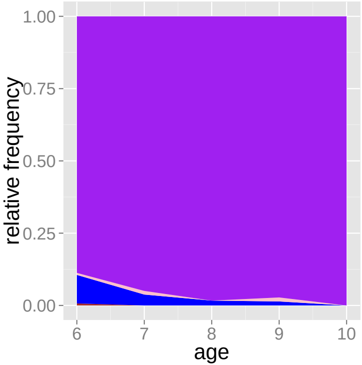
- BLACK, WHITE, RED, and BLUE: hardly any errors;
- GREEN and YELLOW: a few early errors;
- ORANGE: somewhat haphazard, persistent errors;
- PURPLE: persistent errors, mostly *blue* (but not *purple* for BLUE!)

- BLACK, WHITE, RED, and BLUE: hardly any errors;
  - → Predicted correctly by model
- GREEN and YELLOW: a few early errors;
  - → Underestimation of the few errors
- ORANGE: somewhat haphazard, persistent errors;
- PURPLE: persistent errors, mostly *blue* (but not *purple* for BLUE!)

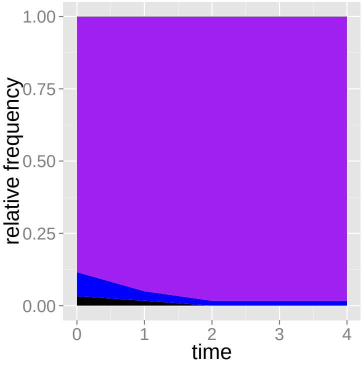
- BLACK, WHITE, RED, and BLUE: hardly any errors;
  $\rightarrow$ Predicted correctly by model
- GREEN and YELLOW: a few early errors;
  $\rightarrow$ Underestimation of the few errors
- ORANGE: somewhat haphazard, persistent errors;
  $\rightarrow$ Underestimation of the errors
- PURPLE: persistent errors, mostly *blue* (but not *purple* for BLUE!)

- BLACK, WHITE, RED, and BLUE: hardly any errors;
  → Predicted correctly by model
- GREEN and YELLOW: a few early errors;
  → Underestimation of the few errors
- ORANGE: somewhat haphazard, persistent errors;
  → Underestimation of the errors
- PURPLE: persistent errors, mostly *blue* (but not *purple* for BLUE!)
  → see following slides . . .

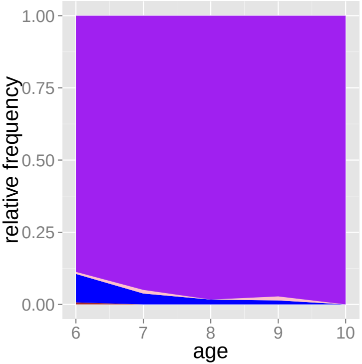
Best fit for PURPLE!



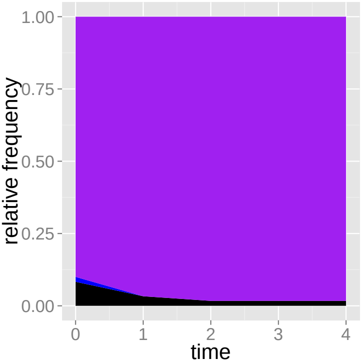
: In Bateman (1915)



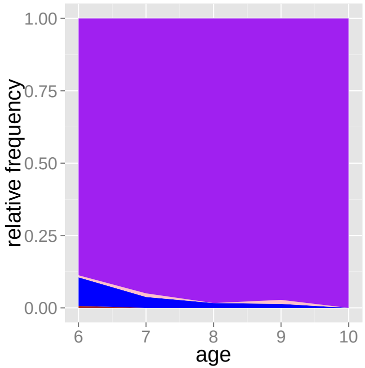
: `perc, relative, gcm`



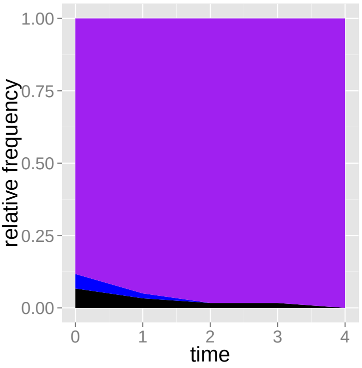
: In Bateman (1915)



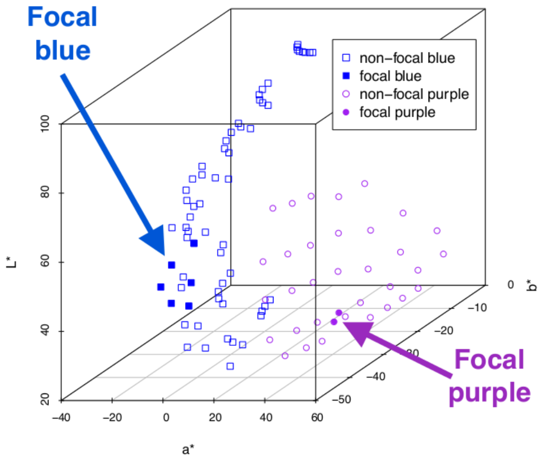
: `perc&conc, relative, gcm`



: In Bateman (1915)



: `perc`, `uniform`, `gcm`


Focal blue
non–focal blue
focal blue
non–focal purple
focal purple
Focal purple
L*
100
80
60
40
20
–40
–20
0
20
40
60
a*
b*
0
–10
–20
–40
–50

- Conceptual dimensions don’t increase fit over perceptual

- Conceptual dimensions don't increase fit over perceptual
- Reason: correlation with perceptual dimensions

	L^*	a^*	b^*
PCA1	-0.01	0.80^*	-0.01
PCA2	-0.97^{***}	0.40	-0.08
PCA3	0.16	-0.03	-0.88^{**}
PCA4	0.60	-0.86^*	0.70

- Conceptual dimensions don't increase fit over perceptual
- Reason: correlation with perceptual dimensions

	L^*	a^*	b^*
PCA1	-0.01	0.80^*	-0.01
PCA2	-0.97^{***}	0.40	-0.08
PCA3	0.16	-0.03	-0.88^{**}
PCA4	0.60	-0.86^*	0.70

- However: why do they correlate strongly but perform much worse independently?

- Role of typological prevalence (vs. perceptual effects and word frequency) in color term acquisition.

- Role of typological prevalence (vs. perceptual effects and word frequency) in color term acquisition.
- Main results:
  - Perceptual features predicted error pattern best.

- Role of typological prevalence (vs. perceptual effects and word frequency) in color term acquisition.
- Main results:
  - Perceptual features predicted error pattern best.

- Role of typological prevalence (vs. perceptual effects and word frequency) in color term acquisition.
- Main results:
  - Perceptual features predicted error pattern best.
  - Typological prevalence (‘conceptual features’) added no error-reduction and performed much worse without perceptual features.

- Role of typological prevalence (vs. perceptual effects and word frequency) in color term acquisition.
- Main results:
  - Perceptual features predicted error pattern best.
  - Typological prevalence ('conceptual features') added no error-reduction and performed much worse without perceptual features.
  - Frequency matters for some colors (see PURPLE)

- Extend this approach to developmental data on more languages and younger children.
- Issue of model behaving too well (underestimating errors).

Thank you!