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

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th>Dutch</th>
<th></th>
<th></th>
<th>Tiriyo</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>on</td>
<td>in</td>
<td>op</td>
<td>aan</td>
<td>om</td>
<td>in</td>
<td>tao</td>
</tr>
<tr>
<td>apple in bowl</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ring on finger</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pen on table</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>painting on wall</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>ring on finger</th>
<th>pen on table</th>
<th>painting on wall</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple in bowl</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ring on finger</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>pen on table</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

- 'apple in bowl'
- 'pen on table'
- 'painting on wall'
- 'ring on finger'
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 → part of meaning space,
  - typological prevalence → part of meaning space,
  - frequency → 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.
Perceptual features

- 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

<table>
<thead>
<tr>
<th></th>
<th>white</th>
<th>pink</th>
<th>orange</th>
<th>...</th>
<th>purple</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip A1</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>chip A2</td>
<td>2</td>
<td>13</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chip I40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>15</td>
</tr>
</tbody>
</table>
Conceptual features

- 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,

<table>
<thead>
<tr>
<th></th>
<th>chip B1</th>
<th>chip C1</th>
<th>chip D1</th>
<th>...</th>
<th>chip I40</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip A1</td>
<td>0.81</td>
<td>0.87</td>
<td>0.98</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>chip B1</td>
<td>0.26</td>
<td>0.42</td>
<td></td>
<td>...</td>
<td>0.96</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chip H40</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>0.81</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>chip B1</th>
<th>chip C1</th>
<th>chip D1</th>
<th>...</th>
<th>chip I40</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip A1</td>
<td>120.4</td>
<td>122.1</td>
<td>136.8</td>
<td>...</td>
<td>142.0</td>
</tr>
<tr>
<td>chip B1</td>
<td></td>
<td>73.6</td>
<td>82.1</td>
<td>...</td>
<td>128.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chip H40</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>112.6</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>PCA1</th>
<th>PCA2</th>
<th>PCA3</th>
<th>...</th>
<th>PCA330</th>
</tr>
</thead>
<tbody>
<tr>
<td>chip A1</td>
<td>2.4</td>
<td>-4.2</td>
<td>3.8</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>chip B1</td>
<td>2.7</td>
<td>-1.9</td>
<td>1.0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chip H40</td>
<td>-4.2</td>
<td>2.2</td>
<td>3.2</td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>
Conceptual features

- 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 $+\text{conc}$, 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.
Train learning model on iteratively sampled $t, s$ pairs

<table>
<thead>
<tr>
<th>$t$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>52.4, 0.8, 0.2, 0.83, 0.23, 0.41, 0.03</td>
</tr>
<tr>
<td>perc</td>
<td>conc</td>
</tr>
</tbody>
</table>

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
The learning models

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
Experimental set-up

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

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

- 30 simulations: distribution over color terms
- $n$ children: distribution over color terms
- take Euclidean distance between them (error).
Comparison 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
Global fit and effect of parameters

<table>
<thead>
<tr>
<th>parameter</th>
<th>setting</th>
<th>mean error</th>
</tr>
</thead>
<tbody>
<tr>
<td>features ***</td>
<td>perc &amp; conc</td>
<td>$\mu = 0.015$</td>
</tr>
<tr>
<td></td>
<td>perc</td>
<td>$\mu = 0.020$</td>
</tr>
<tr>
<td></td>
<td>conc</td>
<td>$\mu = 0.354$</td>
</tr>
<tr>
<td>frequency</td>
<td>relative</td>
<td>$\mu = 0.130$</td>
</tr>
<tr>
<td></td>
<td>uniform</td>
<td>$\mu = 0.130$</td>
</tr>
<tr>
<td>model *</td>
<td>GCM</td>
<td>$\mu = 0.120$</td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>$\mu = 0.139$</td>
</tr>
</tbody>
</table>

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
Findings per color

- **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!*)

See following slides...
Findings per color

- **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**)!
Findings per color

- **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!)
Findings per color

- **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
Findings per color

In Bateman (1915)

Beekhuizen & Stevenson
Perceptual, Conceptual, and Frequency Effects on Error Patterns in English Color Term Acquisition
Findings per color

In Bateman (1915)

: perc, uniform, gcm
Perceptual, Conceptual, and Frequency Effects on Error Patterns in English Color Term Acquisition

Beekhuizen & Stevenson

Department of Computer Science, University of Toronto
Conceptual dimensions don’t increase fit over perceptual
The role of conceptual features

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

<table>
<thead>
<tr>
<th></th>
<th>L*</th>
<th>a*</th>
<th>b*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA1</td>
<td>−0.01</td>
<td>0.80*</td>
<td>−0.01</td>
</tr>
<tr>
<td>PCA2</td>
<td>−0.97***</td>
<td>0.40</td>
<td>−0.08</td>
</tr>
<tr>
<td>PCA3</td>
<td>0.16</td>
<td>−0.03</td>
<td>−0.88**</td>
</tr>
<tr>
<td>PCA4</td>
<td>0.60</td>
<td>−0.86*</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Conceptual dimensions don’t increase fit over perceptual

Reason: correlation with perceptual dimensions

<table>
<thead>
<tr>
<th>PCA</th>
<th>L*</th>
<th>a*</th>
<th>b*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA1</td>
<td>-0.01</td>
<td>0.80*</td>
<td>-0.01</td>
</tr>
<tr>
<td>PCA2</td>
<td>-0.97***</td>
<td>0.40</td>
<td>-0.08</td>
</tr>
<tr>
<td>PCA3</td>
<td>0.16</td>
<td>-0.03</td>
<td>-0.88**</td>
</tr>
<tr>
<td>PCA4</td>
<td>0.60</td>
<td>-0.86*</td>
<td>0.70</td>
</tr>
</tbody>
</table>

However: why do they correlate strongly but perform much worse independently?
What we learned

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

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

- 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.
What we learned

- 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)
What’s there to do

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