Probabilistic models of language acquisition

Klinton Bicknell & Roger Levy

ESSLLI 26
University of Tübingen
22 August 2014
Situating language acquisition

day 1: sound categorization

Figure 5.2: Likelihood functions for /b/–/p/ phoneme categorizations, with µ\(_b\) = 0, µ\(_p\) = 50, σ\(_b\) = σ\(_p\) = 12. Intuitively, the phoneme that is more likely to be realized with VOT in the vicinity of a given input is a better choice for the input, and the greater the discrepancy in the likelihoods the stronger the categorization preference. An input with no negligible likelihood for each phoneme is close to the “categorization boundary”, but may still have a preference. These intuitions are formally realized in Bayes’ Rule:

\[
P(b|\mathbf{x}) = \frac{P(\mathbf{x}|b)P(b)}{P(\mathbf{x})}
\]

and since we are considering only two alternatives, the marginal likelihood is simply the weighted sum of the likelihoods under the two phonemes:

\[
P(\mathbf{x}) = P(\mathbf{x}|b)P(b) + P(\mathbf{x}|p)P(p).
\]

If we plug in the normal probability density function we get

\[
P(b|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x} - \mu_b)^2}{2\sigma^2}} P(b)
\]

and

\[
P(p|\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x} - \mu_p)^2}{2\sigma^2}} P(p)
\]

In the special case where σ\(_b\) = σ\(_p\) we can simplify this considerably by cancelling the
day 1: inferring sound category $c$ from sound token $S$

- $c \sim$ discrete choice, e.g., $p(p) = p(b) = 0.5$
- $S|c \sim \text{Gaussian}(\mu_c, \sigma^2_c)$
Statistical Model

day 2a: sound similarity

Actual Stimulus

Perceived Stimulus
Statistical Model

day 2a: inferring target production $T$ from sound token $S$

Choose a category $c$ with probability $p(c)$

Articulate a target production $T$ with probability $p(T|c)$

$$p(T|c) = N(\mu_c, \sigma_c^2)$$

Listener hears speech sound $S$ with probability $p(S|T)$

$$p(S|T) = N(T, \sigma_S^2)$$
Statistical Model

day 2b: incremental parsing

\[
\begin{array}{c}
S \\
NP \\
NP \quad VP \\
DT \quad \text{NN} \quad V \\
\text{the horse raced} \quad \ldots
\end{array}
\]
Statistical Model

day 2b: inferring syntactic structure $T$ from words $w$

$T \sim \text{PCFG}$

words are leaves of the trees
Statistical Model

day 3: sentence processing and eye movements in reading

The coach smiled at the player tossed the frisbee
The coach smiled at the player thrown the frisbee
The coach smiled toward the player tossed the frisbee
The coach smiled toward the player thrown the frisbee
day 3: inferring words $w$ and trees $T$ from perceptual input $I$

- a tree $T \sim \text{PCFG}$
- words are leaves of the tree
- visual input $I \sim \text{noise}(w)$
day 4: sentence meaning judgments

The zarg prolted the cherid to a really gromious flig.

Which is more likely?
○ The cherid is in a new place. LOCATIVE inference
○ The cherid has a new owner. POSSESSIVE inference
Situating language acquisition

day 4: inferring meaning $M$ from linguistic form $F$ and processing pressures $P$

$F \sim \text{function}(M, P)$
Situating language acquisition

in all these cases

• step 1: identify the relevant sources of information

• step 2: make a generative model in which
  • the thing to be inferred was a latent variable
  • the relevant information was used to specify a prior for the latent variable or the likelihood of the data given that latent variable

• step 3: apply Bayesian inference (relatively easy here: given prior and likelihood)
Problems in acquisition

step 1

let's apply step 1 (identify relevant information sources) to problems in acquisition
Problems in acquisition

learning to segment words

[yuwanttusidəbuk]?  

->  

you want to see the book?

[e.g., Goldwater et al., 2009]
Problems in acquisition

learning word meanings

[e.g., Frank et al., 2009]
Problems in acquisition

Learning phonotactics

[blɪk]

[mdwɨ] Polish mdły: 'tasteless'

[e.g., Hayes & Wilson, 2008]
Problems in acquisition

learning syntax

The boy is hungry. -> Is the boy hungry?

The boy who is smiling is happy. -> ???
  Is the boy who is smiling happy?
  *Is the boy who smiling is happy?

[e.g., Perfors et al., 2011]
Problems in acquisition

learning the sound categories in the language from the sounds

- are long VOTs [t] and short VOTs [d] functionally different sounds or just natural variation?

[e.g., Pajak et al., 2013]
Problems in acquisition

steps 2 and 3: construct a generative model, perform inference

• a bit different than before

• we'll go in depth through an example of sound category learning

• note: many of my visualizations were made by Dr. Bozena Pajak (University of Rochester, Brain & Cognitive Sciences. soon to be my colleague at Northwestern University)
Distributional information

learning sound categories

• how do we learn that [t] and [d] are different categories?
• information source 1: distributional information
experimental evidence

- we know that babies and adults use distributional information to help infer category structure [Maye & Gerken, 2000; Maye et al., 2002]

- Pajak & Levy (2011) performed an experiment replicating this with adults, which I'll describe
Experimental data (Pajak & Levy 2011)

- Distributional training:
  - adult English native speakers exposed to words in a new language, where the middle consonant varied along the length dimension

[aja]_{145\text{ms}}
[ina]_{205\text{ms}}
[ila]_{115\text{ms}}
[ama]_{160\text{ms}}
...
Experimental data (Pajak & Levy 2011)

![Graph showing familiarity frequency across stimuli length continuum (in msec). The graph has two curves: one for bimodal and another for unimodal. The x-axis represents stimuli length continuum in milliseconds, ranging from 100 to 280. The y-axis represents familiarity frequency, ranging from 0 to 16.]
Testing:
- participants made judgments about pairs of words

**Example:** [ama]-[amma]

“Are these two different words in this language or two repetitions of the same word?”

- *dependent measure*: proportion of ‘different’ responses (as opposed to ‘same’) on ‘different’ trials
- if learning is successful, we expect:

<table>
<thead>
<tr>
<th>‘DIFFERENT’ RESPONSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bimodal training</td>
</tr>
<tr>
<td>&gt;</td>
</tr>
<tr>
<td>Unimodal training</td>
</tr>
</tbody>
</table>
Experimental data (Pajak & Levy 2011)
Experimental data (Pajak & Levy 2011)

Proportion of 'different' responses

- Bimodal
- Unimodal
Distributional information

distributional information is used

- adults and babies use distributional information to infer categories
- next steps: put this information into a generative model and perform inference
A generative model

Our model from before of where sounds come from

- \( c \sim \text{discrete choice, e.g., } p(p) = p(b) = 0.5 \)
- \( S|c \sim \text{Gaussian}(\mu_c, \sigma^2_c) \)

Still pretty appropriate!
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]  
\[ i \sim \text{discrete}(\phi) \]  
\[ n = \# \text{observations} \]  
\[ m = \# \text{categories} \]  
\[ \phi \sim \text{distribution()} \]  
\[ \mu_i \sim \text{distribution()} \]  
\[ \Sigma_i \sim \text{distribution()} \]  

'mixture of Gaussians'

set in advance

made up!
Inference

**final step: inference**

- now we have generative model with prior on variables of interest and well-defined likelihood
- but inference is usually quite hard
- two common classes of methods for inference, which I'll describe now
Inference

inference method 1: Expectation Maximization (EM)

• if we knew the category parameters (means, variance, probability), it would be easy to categorize datapoints like before

• if we knew how the datapoints were categorized, it would be easy to find category parameters
  • Q: how?

• idea behind EM:
  • randomly initialize category assignments of datapoints
  • estimate category parameters from current assignments
  • now assign datapoints to categories based on category parameters
  • estimate category parameters from current assignments
  • …
EM illustrations (on board)

- bimodal data
- two categories
Inference

EM issues

• only finds most likely value for variables, not posterior distribution on them

• can get stuck in 'local optima'
Inference

inference method 2: Markov chain Monte Carlo

• like EM, the model is in some state at each point in time (with current guesses for latent variables)

• like EM, the model transitions between states

• unlike EM, the model sometimes moves to lower probability states

• if you calculate the transition probabilities correctly, the amount of time the model spends in each state is proportional to its posterior probability

• thus: don't just get most likely variables out, but full posterior distribution

• also: guaranteed to converge (not getting stuck in local optima)
  • however, this guarantee is only with infinite time…
mixture of Gaussian models to learn phonetic categories

- many groups have had success, at least on simple problems (Vallabha et al., 2007; McMurray et al., 2009; Feldman et al., 2009)
- but it seems clear that other information sources are needed too, because categories overlap too much
- next up: work by Pajak and colleagues on another information source
Modeling category learning

[slides on Pajak, Bicknell, & Levy, 2013]
What may help solve this problem

- Learning may be facilitated by languages’ extensive reuse of a set of phonetic dimensions (Clements 2003)

<table>
<thead>
<tr>
<th>BULGARIAN consonants</th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental/Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>hard</td>
<td>m</td>
<td>(m̥)</td>
<td>n</td>
<td>(n̥)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>mʲ</td>
<td></td>
<td></td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>Plosive</td>
<td>hard</td>
<td>p</td>
<td>b</td>
<td>t d</td>
<td>k g</td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>pʲ</td>
<td>bʲ</td>
<td>tʲ dʲ</td>
<td>c f</td>
<td></td>
</tr>
<tr>
<td>Affricate</td>
<td>hard</td>
<td>ts</td>
<td>dz</td>
<td>tʃ dʒ</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>tsʲ</td>
<td>dzʲ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>hard</td>
<td>f v</td>
<td>s z</td>
<td></td>
<td>x (γ)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>fʲ vʲ</td>
<td>sʲ zʲ</td>
<td></td>
<td></td>
<td>(xʲ)</td>
</tr>
<tr>
<td>Trill</td>
<td>hard</td>
<td>r</td>
<td></td>
<td>rʲ</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>hard</td>
<td>(w)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td>hard</td>
<td>t (l)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td></td>
<td>ḳ</td>
<td></td>
</tr>
</tbody>
</table>

(Wikipedia: Bulgarian phonology)
What may help solve this problem

- Learning may be facilitated by languages’ extensive reuse of a set of phonetic dimensions (Clements 2003)

<table>
<thead>
<tr>
<th>THAI vowels</th>
<th>Front</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unrounded</td>
<td>unrounded</td>
</tr>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Close</td>
<td>/i/</td>
<td>/ι:/</td>
</tr>
<tr>
<td>Close-mid</td>
<td>/ɛ/</td>
<td>/ɛː/</td>
</tr>
<tr>
<td>Open-mid</td>
<td>/æ/</td>
<td>/æː/</td>
</tr>
<tr>
<td>Open</td>
<td>/a/</td>
<td>/aː/</td>
</tr>
</tbody>
</table>

(Locations: Thai language)

LENGTH
Learning may be facilitated by languages’ extensive re-use of a set of phonetic dimensions (Clements 2003)

Existing experimental evidence supports this view
- both infants and adults generalize newly learned phonetic category distinctions to untrained sounds along the same dimension (McClaskey et al. 1983, Maye et al. 2008, Perfors & Dunbar 2010, Pajak & Levy 2011)
How do people learn phonetic categories?

- What are the mechanisms underlying generalization?
- How do learners make use of information about some phonetic categories when learning other categories?
How do people learn phonetic categories?

- Pajak et al.’s proposal:
  - in addition to learning specific categories, people also learn **category types**
Experimental data illustrating generalization across analogous distinctions (from Pajak & Levy 2011)

Our computational proposal for how this kind of generalization might be accomplished

Simulations of Pajak & Levy’s data
Experimental data (Pajak & Levy 2011)

sound class: FRICATIVES

length /s/ /ss/

singleton geminate

sound class: SONORANTS

length /n/ /nn/
Experimental data (Pajak & Levy 2011)

- Distributional training:
  - adult English native speakers exposed to words in a new language, where the middle consonant varied along the length dimension

  - [aja]$_{145\text{ms}}$
  - [ina]$_{205\text{ms}}$
  - [ila]$_{115\text{ms}}$
  - [ama]$_{160\text{ms}}$
  - ...

...
Experimental data (Pajak & Levy 2011)

Stimuli length continuum (in msec)

- Expt1: sonorants
  - [n]-...-[nn]
  - [m]-...-[mm]
  - [l]-...-[ll]
  - [j]-...-[jj]

- Expt2: fricatives
  - [s]-...-[ss]
  - [f]-...-[ff]
  - [ʃ]-...-[ʃʃ]
  - [θ]-...-[θθ]

Familiarization frequency

This difference reflects natural distributions of length in different sound classes.
Testing:
- participants made judgments about pairs of words
  
  **Example: [ama]-[amma]**
  
  “Are these two different words in this language or two repetitions of the same word?”

- dependent measure: proportion of ‘different’ responses (as opposed to ‘same’) on ‘different’ trials

- if learning is successful, we expect:
  
  `Bimodal training > Unimodal training`

- to assess generalization, testing included both trained and untrained sound classes (i.e., both sonorants and fricatives)
Experimental data (Pajak & Levy 2011)

Testing

Stimuli length continuum (in msec)

Familiarization frequency

100 120 140 160 180 200 220 240 260 280

[n]-...-[nn]
[m]-...-[mm]
[l]-...-[ll]
[j]-...-[jj]

EXPT 1: ALIGNED CATEGORIES

EXPT 2: MISALIGNED CATEGORIES

Trained sound class (e.g., ama-amma)
Untrained sound class (e.g., asa-assa)
(e.g., asa-assa)

TesYng

Experimental data (Pajak & Levy 2011)
Experimental data (Pajak & Levy 2011)

EXPT 1: ALIGNED CATEGORIES

- trained
- untrained

EXPT 2: MISALIGNED CATEGORIES

- trained
- untrained

Proportion of ‘different’ responses

Expt1–trained
Expt1–untrained
Expt2–trained
Expt2–untrained
Computational modeling

① How can we account for distributional learning?

② How can we account for generalization across sound classes?
Modeling phonetic category learning

- Mixture of Gaussians approach

\[ d_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \]

phonetic category

data point (perceptual token)

\[ \mu_{s}, \sigma_{s}^2 \]
\[ \mu_{ss}, \sigma_{ss}^2 \]

length /s/ /ss/

120ms 201ms 165ms 182ms 115ms ...
Our general approach (following Feldman et al. 2009):

- learning via nonparametric Bayesian inference
- using Dirichlet processes, which allow the model to learn the number of categories from the data

\[ G_0 \]

\[ z_i \]

\[ d_i \]

\[ i \in \{1..n\} \]
Computational modeling

① How can we account for distributional learning? ✓

② How can we account for generalization across sound classes?
In addition to acquiring specific categories, learners infer **category types**, which can be shared across sound classes.

This means that already learned categories can be directly re-used to categorize other sounds.

To implement this proposal, we use a **hierarchical Dirichlet process**, which allows for sharing categories across data groups (here, sound classes).
Modeling generalization: HDP

prior

phonetic category

datapoint (perceptual token)

c = sonorants

\[ z_{ic} \sim G_c \]
\[ d_{ic} \sim \mathcal{N}(\mu_{z_{ic}}, \sigma_{z_{ic}}^2) \]

son-sg son-gem

sonorants

length /n/ /nn/

c \in C

\[ i \in \{1..n_c\} \]
But people are able to generalize even when analogous category types are implemented phonetically in different ways.

We want the model to account for potential differences between sound classes.
Modeling generalization: HDP

Accounting for differences between sound classes:

learnable class-specific ‘offsets’ by which data in a class are shifted along a phonetic dimension (cf. Dillon et al. 2013)

\[
H : \quad \mu \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0}) \\
\sigma^2 \sim \text{InvChiSq}(\nu_0, \sigma^2_0) \\
\]

\[
G_0 \sim DP(\gamma, H) \\
G_c \sim DP(\alpha_0, G_0) \\
z_{ic} \sim G_c \\
f_c \sim \mathcal{N}(0, \sigma^2_f) \\
d_{ic} \sim \mathcal{N}(\mu_{z_{ic}}, \sigma^2_{z_{ic}}) + f_c
\]

fricatives

length /s/ /ss/

sonorants

length /n/ /nn/
Simulation results

EXPT 1: ALIGNED CATEGORIES

EXPT 2: MISALIGNED CATEGORIES

HUMAN DATA
Simulation results: NO OFFSET PARAMETER

**EXPT 1: ALIGNED CATEGORIES**

- Trained: Bimodal
- Untrained: Bimodal

**EXPT 2: MISALIGNED CATEGORIES**

- Trained: Bimodal
- Untrained: Bimodal

---

**HUMAN DATA**

- Proportion of 2-category inferences
  - Expt1-train: Bimodal
  - Expt1-unt: Unimodal
  - Expt2-train: Bimodal
  - Expt2-unt: Unimodal

---

`we need the shift parameter to account for generalization across misaligned categories!`
Modeling category learning

Other work on category learning

- Feldman et al. (2009) add a (latent) lexicon to category learning
Conclusions

probabilistic models in acquisition

• in general: still working on trying to incorporate many sources of information

• it seems that infants, children, and adults use many sources of information to learn language

• one very exciting information source: other languages
Conclusions

computational psycholinguistics

- this course gave a very broad overview of many areas: perception, comprehension, production, acquisition
- gave a sense for how probabilistic models can be used
- also covered much of the essential technical knowledge to use probabilistic models
- this is still very much the beginning of probabilistic modeling in the study of language: very many open questions and models yet to build!
- hope you enjoyed!