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. For the input
\(x = 27\), the likelihoods favor /p/.

Figure 5.3: Posterior probability curve for Bayesian phoneme discrimination as
a function of VOT

We further simplify the problem by characterizing any acoustic representation
\(x\) as a single real-valued number representing the VOT, and the
likelihood functions for /b/ and /p/ as normal density functions (Section 2.10) with means
\(µ\(_b\)\), \(µ\(_p\)\) and standard deviations
\(σ\(_b\)\), \(σ\(_p\)\) respectively.

Figure 5.2 illustrates the likelihood functions for the choices
\(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:

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

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

\[
P(x) = P(x|/b)P(/b) + P(x|/p)P(/p).
\]

If we plug in the normal probability density function we get

\[
P(/b| x) = \frac{1}{\sqrt{2\pi\sigma^2_b}}\exp\left(-\frac{(x - µ_b)^2}{2\sigma^2_b}\right)P(/b)
\]

\[
P(/p| x) = \frac{1}{\sqrt{2\pi\sigma^2_p}}\exp\left(-\frac{(x - µ_p)^2}{2\sigma^2_p}\right)P(/p)
\]
Situating language acquisition

day 1: inferring sound category $c$ from sound token $S$
Situating language acquisition

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 $S$
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

```
S
  | NP
  |   NP  VP
  |    |   |
  |    DT NN V
  |      |
  |      the horse raced ...
```
Statistical Model

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

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 $I \sim \text{noise}(w)$
Situating language acquisition

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

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

in all these cases
Situating language acquisition

in all these cases

• step 1: identify the relevant sources of information
Situating language acquisition

in all these cases

- step 1: identify the relevant sources of information
- step 2: make a generative model in which
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
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
Situationing 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

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

step 1

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

[yuwanttusiðəbʊk]?  

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

learning to segment words

[yuwantttusuðəbʊk]?  

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

learning to segment words

[yuwanttusidębuk]? 

->

you want to see the book?

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

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

learning word meanings

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

[blink]

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

learning phonotactics

[blɪk]

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

learning phonotactics

[blɪk]

[mdwɪ]

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

learning phonotactics

[blɪk]

[mdwɨ] Polish mdły: 'tasteless'

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

The boy is hungry. \(\rightarrow\) Is the boy hungry?

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

learning syntax

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

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

learning syntax

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

The boy who is smiling is happy. -> ???

[e.g., Perfors et al., 2011]
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?

[e.g., Perfors et al., 2011]
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

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

learning the sound categories in the language from the sounds

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

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

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

• a bit different than before
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
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
Distributional information

learning sound categories
Distributional information

learning sound categories

- how do we learn that [t] and [d] are different categories?
Distributional information

learning sound categories

- how do we learn that [t] and [d] are different categories?
- information source 1: distributional information
learning sound categories

- how do we learn that [t] and [d] are different categories?
- information source 1: distributional information
learning sound categories

- how do we learn that [t] and [d] are different categories?
- information source 1: distributional information
Distributional information
Distributional information

experimental evidence
experimental evidence

- we know that babies and adults use distributional information to help infer category structure [Maye & Gerken, 2000; Maye et al., 2002]
Distibutional 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 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]_{145ms}  [ina]_{205ms}  [ila]_{115ms}  [ama]_{160ms}  ...


Experimental data (Pajak & Levy 2011)
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?”

Experimental data (Pajak & Levy 2011)
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
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:

  ‘DIFFERENT’ RESPONSES

- Bimodal training > Unimodal training

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

Stimuli length continuum (in msec)

- **Familiarization frequency**
  - **bimodal**
  - **unimodal**

**Expt1:** sonorants

**Expt2:** fricatives

**Test stimuli**
Experimental data (Pajak & Levy 2011)

Proportion of 'different' responses

- Bimodal
- Unimodal
Distributional information
Distributional information is used
Distributional information is used

- adults and babies use distributional information to infer categories
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
A generative model

our model from before of where sounds come from
A generative model

our model from before of where sounds come from

• $c \sim$ discrete choice, e.g., $p(p) = p(b) = 0.5$

• $S|c \sim \text{Gaussian}(\mu_c, \sigma^2_c)$
A generative model

our model from before of where sounds come from

- $c \sim$ 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

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

our model from before of where sounds come from

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

still pretty appropriate!
A generative model

our model from before of where sounds come from

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

still pretty appropriate!
A generative model
A generative model

a model of where categories and sounds come from

Roger Levy – Probabilistic Models in the Study of Language
draft, November 6, 2012
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]
A generative model

a model of where categories and sounds come from

\[
\phi \rightarrow i \rightarrow y \quad \text{(S before)}
\]

\[
\Sigma \quad \mu \quad n \quad m
\]

(S before) \( y \sim N(\mu_i, \Sigma_i) \)

(\( \sigma^2 \) before)
A generative model

a model of where categories and sounds come from

(S before) \( y \sim N(\mu_i, \Sigma_i) \)

(\( \sigma^2 \) before) \( i \sim \text{discrete}(\phi) \)
A generative model

a model of where categories and sounds come from

\[
\begin{align*}
\phi & \xrightarrow{} i \\
\Sigma & \xrightarrow{} \mu \\
& \xrightarrow{} y \\
& \xrightarrow{} i \\
& \xrightarrow{} \Sigma
\end{align*}
\]

\(y \sim N(\mu_i, \Sigma_i)\) 

\((\sigma^2 \text{ before})\) 

\((S \text{ before})\) 

\((c \text{ before})\)

\(i \sim \text{discrete}(\phi)\)
A generative model

a model of where categories and sounds come from

\[ (\sigma^2 \text{ before}) \]

(S before) \( y \sim N(\mu_i, \Sigma_i) \)

(c before) \( i \sim \text{discrete}(\phi) \)

\[ n = \# \text{ observations} \]
A generative model

A model of where categories and sounds come from

(S before) \( y \sim N(\mu_i, \Sigma_i) \)

(c before) \( i \sim \text{discrete} (\phi) \)

\( n = \# \text{ observations} \)

\( m = \# \text{ categories} \)
A generative model

A model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]

(c before) \[ i \sim \text{discrete}(\phi) \]

\[ n = \text{# observations} \]

\[ m = \text{# categories} \]
A generative model

a model of where categories and sounds come from

(S before) $y \sim N(\mu_i, \Sigma_i)$

(c before) $i \sim \text{discrete}(\phi)$

$n = \# \text{observations}$

$m = \# \text{categories}$

$\phi \sim \text{distribution()}$
A generative model

a model of where categories and sounds come from

\[
\begin{align*}
\phi \\
i \\
y \\
\Sigma \\
\mu \\
m
\end{align*}
\]  

\[
\begin{align*}
(S \text{ before}) & \quad y \sim N(\mu_i, \Sigma_i) \\
(c \text{ before}) & \quad i \sim \text{discrete}(\phi) \\
\end{align*}
\]

\[
\begin{align*}
n = \# \text{ observations} \\
m = \# \text{ categories} \\
\phi \sim \text{distribution()} \\
\mu_i \sim \text{distribution()}
\end{align*}
\]
A generative model

a model of where categories and sounds come from

(S before) \( y \sim N(\mu_i, \Sigma_i) \)

(c before) \( i \sim \text{discrete}(\phi) \)

\( n = \# \text{observations} \)

\( m = \# \text{categories} \)

\( \phi \sim \text{distribution()} \)

\( \mu_i \sim \text{distribution()} \)

\( \Sigma_i \sim \text{distribution()} \)
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]

(S before)

(c before) \[ i \sim \text{discrete}(\phi) \]

\[ n = \# \text{observations} \]

\[ m = \# \text{categories} \]

\[ \phi \sim \text{distribution()} \]

\[ \mu_i \sim \text{distribution()} \]

\[ \Sigma_i \sim \text{distribution()} \]

'mixture of Gaussians'
A generative model

A model of where categories and sounds come from

(S before) \( y \sim \mathcal{N}(\mu_i, \Sigma_i) \)

(c before) \( i \sim \text{discrete}(\phi) \)

\( n = \# \text{observations} \)

\( m = \# \text{categories} \)

\( \phi \sim \text{distribution()} \)

\( \mu_i \sim \text{distribution()} \)

\( \Sigma_i \sim \text{distribution()} \)

'mixture of Gaussians'

made up!
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]

(c before) \( i \sim \text{discrete}(\phi) \)

\( n = \# \text{observations} \)

\( m = \# \text{categories} \) set in advance

\[ \phi \sim \text{distribution()} \]

\[ \mu_i \sim \text{distribution()} \]

\[ \Sigma_i \sim \text{distribution()} \] made up!

'mixture of Gaussians'
Inference
final step: inference
Inference

**final step: inference**

- now we have generative model with prior on variables of interest and well-defined likelihood
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
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

inference method 1: Expectation Maximization (EM)
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
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
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?
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:
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
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
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
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
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
  • …
Inference
Inference

EM illustrations (on board)
Inference

EM illustrations (on board)

• bimodal data
Inference

EM illustrations (on board)

- bimodal data
- two categories
Inference
Inference

EM issues
Inference

**EM issues**

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

**EM issues**

- only finds most likely value for variables, not posterior distribution on them
- can get stuck in 'local optima'
Inference

**EM issues**

- only finds most likely value for variables, not posterior distribution on them
- can get stuck in 'local optima'
Inference
Inference

inference method 2: Markov chain Monte Carlo
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)
Inference

inference method 2: Markov chain Monto 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
Inference

inference method 2: Markov chain Monto 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
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
Inference

inference method 2: Markov chain Monto 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
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)
Inference

inference method 2: Markov chain Monto 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…
Modeling category learning
Modeling category learning

mixture of Gaussian models to learn phonetic categories
Modeling category learning

- 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)
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
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
Modeling category learning

[slides on Pajak, Bicknell, & Levy, 2013]
Normally not considered part of the phonetic inventory of the Bulgarian language. The Bulgarian obstruent consonants are divided into 12 pairs voiced vs. voiceless on the criteria of sonority. The only obstruent without a counterpart is the voiceless velar fricative /x/. The contrast 'voiced vs. voiceless' is neutralized in word-final position, where all obstruents are voiceless (as in most Slavic languages); this neutralization is, however, not reflected in the spelling.

Bulgarian consonants

<table>
<thead>
<tr>
<th>Place</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>(ñ)</td>
<td>n</td>
<td>(ñ)</td>
<td>j</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>mʲ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plosive</td>
<td>hard</td>
<td>p</td>
<td>t</td>
<td>d</td>
<td>k</td>
<td>g</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>pʲ</td>
<td>tʲ</td>
<td>dʲ</td>
<td>c</td>
<td>ɟ</td>
</tr>
<tr>
<td>Affricate</td>
<td>hard</td>
<td>tʃ</td>
<td>ts</td>
<td>dz</td>
<td>tʃ</td>
<td>dz</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>tʃʲ</td>
<td>tsʲ</td>
<td>dzʲ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>hard</td>
<td>f</td>
<td>s</td>
<td>z</td>
<td>x</td>
<td>(ɣ)</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>fʲ</td>
<td>sʲ</td>
<td>zʲ</td>
<td></td>
<td>(xʲ)</td>
</tr>
<tr>
<td>Trill</td>
<td>hard</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>rʲ</td>
<td></td>
<td></td>
<td></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>Ɂ (l)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ʎ</td>
</tr>
</tbody>
</table>

^1 According to Klagstad Jr. (1958:46–48), /t tʲ d dʲ s sʲ z zʲ n/ are dental. He also analyzes /ɲ/ as palatalized dental nasal, and provides no information about the place of articulation of /t͡ʃ t͡ʃʲ s s͡z t͡ʃ t͡ʃʲ d dʲ c cʲ/.

^2 Only as an allophone of /m/ and /n/ before /f/ and /v/. Examples: инфлация [iɱˈflaʦijɐ] 'inflation'.

^3 /ɣ/ exists as an allophone of /x/ only at word boundaries before voiced obstruents. Example: видях го [viˈdʲaŋgo] 'I saw him'.

^4 /l/ can be analyzed as an allophone of /ɫ/ as it appears only before front vowels. A trend of l-vocalization is emerging among younger native speakers and more often in colloquial speech.

Learning may be facilitated by languages’ extensive re-use of a set of phonetic dimensions (Clements 2003)
What may help solve this problem

- Learning may be facilitated by languages’ extensive re-use 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>/iː/</td>
</tr>
<tr>
<td>Close-mid</td>
<td>/e/</td>
<td>/eː/</td>
</tr>
<tr>
<td>Open-mid</td>
<td>/ɛ/</td>
<td>/ɛː/</td>
</tr>
<tr>
<td>Open</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Wikipedia: Thai language)
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)
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**

![Diagram showing singleton <-> geminate and generalization]
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] 145ms
  - [ina] 205ms
  - [ila] 115ms
  - [ama] 160ms
  - ...

[Diagram showing time differences and example words]
Experimental data (Pajak & Levy 2011)

Training

This difference reflects natural distributions of length in different sound classes.
Experimental data (Pajak & Levy 2011)

- 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: 'DIFFERENT RESPONSES'

    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)

EXPT 1: ALIGNED CATEGORIES

EXPT 2: MISALIGNED CATEGORIES

Stimuli length continuum (in msec)
Experimental data (Pajak & Levy 2011)

**EXPT 1:** ALIGNED CATEGORIES
- Trained
- Untrained

**EXPT 2:** MISALIGNED CATEGORIES
- Trained
- Untrained

Proportion of 'different' responses

- Bimodal
- Unimodal

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**

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

- phonetic category
- datapoint (perceptual token)

120ms 201ms 165ms 182ms 115ms ...
Modeling phonetic category learning

- **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

```
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
```
Computational modeling

① How can we account for distributional learning? ✓

② How can we account for generalization across sound classes?
Proposal

- 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).

![Diagram showing fricatives and sonorants with length and examples of /s/, /ss/, /n/, /nn/ symbols.]
Modeling generalization: HDP

prior

phonetic category

datapoint (perceptual token)

c = sonorants

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

\[ z_{ic} \sim G_c \]

\[ G_0 \sim \text{DP}(g, H) \]

\[ G_c \sim \text{DP}(a_0, G_0) \]

\[ c = \text{sonorants} \]

fricatives

length /s/ /ss/

sonorants

length /n/ /nn/

frik-sg fric-gem

\[ \mu_{\text{fric-sg}}, \sigma_{\text{fric-sg}}^2 \]

\[ \mu_{\text{fric-gem}}, \sigma_{\text{fric-gem}}^2 \]
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)

\[
H : \mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \\
\sigma^2 \sim \text{InvChiSq}(\nu_0, \sigma_0^2) \\
G_0 \sim \text{DP}(\gamma, H) \\
G_c \sim \text{DP}(\alpha_0, G_0) \\
z_{ic} \sim \mathcal{G}_c \\
f_c \sim \mathcal{N}(0, \sigma_f^2) \\
d_{ic} \sim \mathcal{N}(\mu_{z_{ic}}, \sigma_{z_{ic}}^2) + f_c
\]
Simulation results

**EXPT 1:** ALIGNED CATEGORIES

**EXPT 2:** MISALIGNED CATEGORIES

**HUMAN DATA**

Proportion of 2−category inferences

- **EXPT 1:** ALIGNED CATEGORIES
  - Trained
  - Untrained

- **EXPT 2:** MISALIGNED CATEGORIES
  - Trained
  - Untrained

**Extended model:**

- **EXPT 1:** ALIGNED CATEGORIES
  - Expt1−trained
  - Expt1−untrained

- **EXPT 2:** MISALIGNED CATEGORIES
  - Expt2−trained
  - Expt2−untrained
Simulation results: NO OFFSET PARAMETER

**EXPT 1: ALIGNED CATEGORIES**

- **Proportion of 'different' responses**
  - 0.00 to 1.00
  - Bimodal vs. Unimodal

**EXPT 2: MISALIGNED CATEGORIES**

- **Proportion of 'different' responses**
  - 0.00 to 1.00
  - Bimodal vs. Unimodal

**HUMAN DATA**

- **Proportion of 2-category inferences**
  - Trained vs. Untrained

---

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

Discussion

![Graphs of vowel categories](image-url)
Modeling category learning

other work on category learning

Figure 3: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories (a) computed from all speakers' vowel productions from Hillenbrand et al. (1995) and learned by the (b) lexical-distributional model, (c) distributional model, (d) gradient descent algorithm.

Figure 4: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories (a) computed from all speakers' vowel productions from Hillenbrand et al. (1995) and learned by the (b) lexical-distributional model, (c) distributional model, (d) gradient descent algorithm.
Modeling category learning

other work on category learning

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

![Graphs showing distributional and lexical-distributional models for vowel categories.](image)
Conclusions
Conclusions

probabilistic models in acquisition
Conclusions

probabilistic models in acquisition

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

computational psycholinguistics
Conclusions

computational psycholinguistics

• this course gave a very broad overview of many areas: perception, comprehension, production, acquisition
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
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
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!
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!