An Introduction to Learning

Lecture 11/13

Todd M. Gureckis
Department of Psychology
New York University
todd.gureckis@nyu.edu
The “ivory towers” of learning

Analogical learning, causal learning, and how learning a language might be possible (part I)
Fig. 2. (A) A centrally symmetric convex region shown as centered on 0, as centered on x, and as having a center, c, falling within the intersection of the regions centered on 0 and on x. (B) For an illustrative nonconvex region centered on 0, the locus of centers, c, of similarly shaped regions having a constant (approximately 20%) overlap with the region centered on 0 (dotted curve); and an ellipse corresponding to the Euclidean metric (smooth curve).

Consequential Region
... but

- Is feature similarity enough to explain the “psychological space” of similarity?
Why We’re So Smart

Dedre Gentner

- The ability to draw abstractions from particulars—to generalize experience and store regularities across vastly different cases
- The ability to maintain hierarchies of abstraction, so that we can store information about Fido, about dachshunds, about dogs, or about living things
- The ability to concatenate assertions and arrive at a new conclusion
- The ability to reason outside of the current context—to think about different locations and different times and even to reason hypothetically about different possible worlds
- The ability to compare and contrast two representations to discover where they are consistent and where they differ
- The ability to reason analogically—to notice common relations across different situations and project further inferences
- The ability to learn and use external symbols to represent numerical, spatial, or conceptual information
- The ability to learn symbols that lack any iconic relation to their referents
- The ability to learn and use symbols whose meanings are defined in terms of other learned symbols, including even recursive symbols such as the set of all sets
- The ability to invent and learn terms for abstractions as well as for concrete entities
- The ability to invent and learn terms for relations as well as things
The “career of similarity”
Quine (1960); Genter & Rattermann (1991)

“Brute” perceptual similarity

“Theoretical” similarity

Age/Development
Young infants and concrete, object-based transfer

- 10-month-old infants can learn to pull cloth to get a toy, but fail to transfer to a new situation unless it was highly (perceptually) similar to initial task (Chen, Sanchez, & Campbell, 1997)

- Oakes and Cohen (1990) found evidence of “conservatism” in interpreting causal events (preferred matches that preserved object identity over relational properties)
Overall Perceptual Similarity in Young Kids

Smith (1998)
Structure Mapping Theory

• A theory of relational similarity, analogy, and comparison based on complex, structured information rather than exclusively perceptual features

• Formalized as a graph matching problem (two domains are represented as objects in semantic memory, the target and the base) and the goal is to figure out a correspondence between them

(1) The dog chased the cat.
(2) The coyote chased the lynx.
(3) The shark chased the mackerel.
(4) Amalgamated Tire Co. made a takeover bid for Racine Ironworks.
(5) The cat chased the mouse.
Constraints on Mapping

- **One to one** (one part or object in one the base maps to one object in the target)

- **Parallel connectivity** (if elements correspond across the two representations, then the elements they govern must correspond as well).

- **Systematicity principle** (richer, deeper relational matches are preferred, preferring deeper matches even if number of matches is the same at a lower level)
Figure 7.1. Examples of physical situations involving (a) water flow and (b) heat flow (adapted from Buckley, 1979, pp. 15–25).
Structure Mapping Theory

Figure 7.5. Representations of water and heat given to the structure-mapping engine.
Figure 7.2. A representation of the water situation. Predicates are written in upper case and circled; objects are written in lower case and uncircled.
Structure Mapping Theory

Figure 7.3. A representation of the heat situation that results from the heat/water analogy.
SMT explains...

- Developmental shifts in similarity based processing
- Schema abstraction
- Projection of candidate inferences from one item to another
- Re-representation - altering one or both representations to improve the match
- Promoting attention to relevant differences
- Restructuring - altering one domain in terms of another
Rattermann & Gentner (1991)

Contrasted perceptual matches with relational matches and found that simple objects were better, and that relational language (mommy, daddy, child) helped to promote relational processing.

Suggests that language helps to activate knowledge about relations.

Figure 8.1
Materials and results for the Rattermann and Gentner spatial mapping task with cross-mapped objects, in which object matches compete with relational matches. In the top figures (top), asterisks and the solid line denote the correct match; the dashed line denotes the incorrect object match. The graph (bottom) shows the results for 3-, 4-, and 5-year-old children in the baseline condition and for 3-year-old children given the relational labels daddy, mommy, baby.
Relational processing in other species

- Chimpanzees can’t do a “choose opposite” task without symbol training (can learn to choose a lower valued arabic numeral to select candies, but not to point to lower amount to mean larger amount)... symbolic processing? discrimination?

- Can learn relational match to sample tasks with symbolic training (Oden, Thompson, & Premack, 2001)

- Macaque monkeys given the same training with same and different symbols can’t do the relational task (Washburn, Thompson, & Oden, 1997)
..but relational processing in pigeons?
Figure 1. Trained *same* and *different* displays involving 16 icons distributed within a $4 \times 4$ array.
Figure 2. *Same* and *different* displays involving 16 icons distributed within a $5 \times 5$ array.
Figure 3. Examples of the spatially disarranged same and different arrays that were used in Experiment 1.
• Fixation display, single peck displays screen with icon display

• A number of pecks to the screen hides the icon display and lights up two choice regions. Pecks to one are reinforced with a food pellet delivered at a different place in the box

• In testing saw a majority of novel “same”/”different” displays and a number of disarranged versions
Figure 4. Discriminative performance to the original and the disarranged arrays of Experiment 1.
Young & Wasserman, 2001

- Exp. 2 upset the displays even more.
- Four types: same/different uniform/varied, 4-4-4-4 uniform/varied
- 4-4-4-4 items control for same-varied being seen as four sets of 4 new items instead of just rotations of a single item. If they think of them as four rotations of different items same-varied and 4-4-4-4 should be the same. If not same-varied should be like same-uniform
Young & Wasserman, 2001

Same–Uniform (270°)  

Same–Varied
Young & Wasserman, 2001

Different–Uniform (180°)  Different–Varied
Young & Wasserman, 2001

4-4-4-4—Uniform (90°)

4-4-4-4—Varied
Young & Wasserman, 2001

- Initial testing phase to assess same/different uniform discrimination with uniform rotations
- In “real” training, learn to discrimination same-uniform and different-uniform for all four cardinal rotations
- Tested with same-uniform, different-uniform (at all rotations) and same-varied, different varied, 4-4-4-4 uniform and 4-4-4-4-varied
Young & Wasserman, 2001

- Initial testing showed evidence of learning same/different but also that pigeons detected the rotations

- Importantly in main test, rotations had not effect. In addition, the 4-4-4-4 items (four new items) were not treated the same as the uniform rotations.
Figure 6. Discriminative performance for the *same*, *different*, and 4-4-4-4 arrays of Experiment 2 with both consistent (*uniform*) and inconsistent (*varied*) within-display rotations of the icons.
Language! Tales from the front-line of the battle about learning
The language-game is ... not unreasonable (or unreasonable). It is there -- like our life - Wittgenstein

(Despite rejecting that “game” could be simply defined!!)

How do kids learn to play the game? What information exists in the world about the rules of how to play?
Language: What cues are there?
Language: What cues are there?
Language: What cues are there?
I like to think (and the sooner the better!) of a cybernetic meadow where mammals and computers live together in mutually programming harmony like pure water touching clear sky.

I like to think (right now, please!) of a cybernetic forest filled with pines and electronics where deer stroll peacefully past computers as if they were flowers with spinning blossoms.

I like to think (it has to be!) of a cybernetic ecology where we are free of our labors and joined back to nature, returned to our mammal brothers and sisters, and all watched over by machines of loving grace. - Richard Brautigan
Language: What cues are there?
Language: What cues are there?

DEPARTURE FROM EQUIPROBABILITY
(AKA THE DISTRIBUTIONAL HYPOTHESIS, Harris, 1991)
Language: What cues are there?

- A learner who is “attuned” to the right statistical properties of the world can learn the structure.

- This is a computationally feasible proposal (our simple Hebbian learning network can approximate various correlations between parts of its input. More sophisticated processes are possible.

- **KEY IDEA:**

  STRUCTURE from REGULARITY in EXPERIENCE
Language: What cues are there?

- Record all the input to babies in their early life, see what cues they use to acquire language
- Corpora of infant-directed speech which has been transcribed (MacWhinney, 2000)
- Record video and audio (MIT Speechome Project)
- ... but you shouldn’t need to be convinced that language is clearly non-random
Language Learning: Where do we start?

http://www.youtube.com/watch?v=FQjgsQ5G8ug

Fig. 1. A speech waveform of the sentence “Where are the silences between words?” The height of the bars indicates loudness, and the x-axis is time. This example illustrates the lack of consistent silences between word boundaries in fluent speech. The vertical gray lines represent quiet points in the speech stream, some of which do not correspond to word boundaries. Some sounds are represented twice in the transcription below the waveform because of their continued persistence over time.
Saffran (1996)


- 8 Month Old Infants were tested for their ability to learn a “micro” language composed of 4 words
  - ba-du-bi
  - ba-du-wo
  - fu-ku-ji
  - ji-hu-tu

- Read continuously by a computer for 2 minutes
Saffran (1996)

Train

- tu-pi-ro
- go-la-bu
- bi-da-ku
- pa-do-ti

Test

Words
- tu-pi-ro
- go-la-bu

Non-words
- da-pi-ku
- ti-la-do
Saffran (1996)

Train
- pa-bi-ku
- ti-bu-do
- go-la-tu
- da-ro-pi

Test
Words
- pa-bi-ku
- go-la-tu

Part-words
- tu-da-ro
- pi-go-la
Experiment 1

Looking Time in Sec.

Word  Non-word
6.5  7.5  8.5  9.0

Experiment 2

Looking Time in Sec.

Word  Part-word
6.5  7.5  8.5  9.0

Infants  LASR
Language Learning: Where do we start?

What cues are available?
Social, Perceptual, Statistical, Auditory, Intentional, Embodied, etc..

Figure 1: Multi-camera sensing system. The child and the mother play with a set of toys at a table. Two mini cameras are placed onto the child’s and the mother’s heads respectively to collect visual information from two first-person views. A third camera mounted on the top of the table records the bird-eye view of the whole interaction.

Figure 2: The overview of data processing using computer vision techniques. Left: we first remove background pixels from an image and then spot objects and hands in the image based on pre-trained object models. The visual information from two views is then aligned for further data analyses. Right: the processing results from the bird-eye view camera.
Example: Word Learning

- Kids learn how to use words appropriately given very little positive evidence (a few examples) and little explicitly negative evidence... yet, a few examples of “dog” are enough to set them off correctly naming dog-like objects and not calling cars “dogs”

- A famous philosophical problem (Quine’s riddle of induction)

- The standard story line in the developmental literature is that there are constraints that learner come equipped with that limit the types of generalization they will do in any situation (e.g., taxonomic assumption, mutual exclusivity, bias toward basic level).
Example: Word Learning

(a) Generalization judgments (averaged over all stimulus categories)

Training examples:

1
3 subordinate
3 basic
3 superordinate

Test object match level:
subord. basic superord. nonmatch

... ... ... ...
Example: Word Learning
What is the hypothesis space?

\[ p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')} \]

- each hypothesis is a proposal about which levels of the tree a word applies.
What is the prior?

\[ p(h \mid d) = \frac{p(d \mid h)p(h)}{\sum_{h' \in H} p(d \mid h')p(h')} \]

\[ p(h) \propto \text{height(PARENT}(h)) - \text{height}(h). \]

Wait, but that’s not a probability! Called an “improper prior”, ok because we are normalizing
What is the likelihood?

\[
p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')}
\]

\[1/K^n.\]

\[p(X|h) \propto \left[\frac{1}{\text{height}(h) + \epsilon}\right]^n,\]

Size principal... larger hypothesis assign lower probability to any single event. Thus, more complex hypotheses need more data to support
(b) Marginal likelihood
Predicting the label of a new item, $y$

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X).$$
Examples: 1 3 subordinate 3 basic 3 superordinate

(a) Bayesian model
(b) Bayesian model (including basic-level bias in prior)

(c) Pure similarity model (Bayes minus size principle)
(d) Pure rule model (Bayes minus hypothesis averaging)

Figure 3: Predictions of the basic Bayesian model and three variants for the data in Figure 1.
Example: Word Learning

- What did the model add to our understanding of word learning?

- Show how the interaction between different constraints could be incorporated into a single process

- Emphasized how the structure of the learner’s hypothesis space might serve to limit generalization... indeed, there is a kind of implicit negative feedback in that each time an object is NOT called something it is incorporated into the model’s posterior

- Size principal clearly demonstrates why learning super-ordinates should be harder (have to accumulate evidence from a wide variety of examples

- The model learns FAST (only a few examples)
Marcus et al. (1999)


- 7 Month Old Infants were tested for their ability to learn a “micro” language composed of 16 “sentences” of the form ABA or ABB

ba-du-ba    wo-du-wo    ku-ji-ku    ji-hu-ji

- Read continuously by a computer for 2 minutes
Marcus et al. (1999)
Marcus et al. (1999)

Table 1. Mean time spent looking in the direction of the consistent and inconsistent stimuli in each condition for experiments 1, 2, and 3, and significance tests comparing the listening times. Mean ages of the infants tested were 6 months 27 days (median, 6 months 24 days) in experiment 1, 7 months 1 day (median, 7 months) in experiment 2, and 7 months (median, 7 months 2 days) in experiment 3.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Consistent sentences (s) (SE)</th>
<th>Inconsistent sentences (s) (SE)</th>
<th>Repeated measures analysis of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.3 (0.65)</td>
<td>9.0 (0.54)</td>
<td>$F(14) = 25.7, P &lt; 0.001$</td>
</tr>
<tr>
<td>2</td>
<td>5.6 (0.47)</td>
<td>7.35 (0.68)</td>
<td>$F(14) = 25.6, P &lt; 0.005$</td>
</tr>
<tr>
<td>3</td>
<td>6.4 (0.38)</td>
<td>8.5 (0.5)</td>
<td>$F(14) = 40.3, P &lt; 0.001$</td>
</tr>
</tbody>
</table>
The meaning of complex expressions is determined by the meaning of the parts and the rules used to combine them.

Meaning of “Todd is a Professor” is partly structure “X is a Y” and from the parts “Todd” (me) and “Professor” (a job one can have).

The “rules” of the game are not lookup table rules then, but COMPOSITIONAL SEMANTICS (i.e., systematic structural rules governing the combination of parts).

The meaning of the parts is given by the mental lexicon.
Language: Finding Meaning in the Mental Lexicon

- What does any single concept mean?
- Can look up in a unending dictionary search?
- In other words, can a dictionary describe the human lexicon?
- Raises the grounding problem (Harnad, 1990)... can a real meaning be defined entirely in a structural way without touching perception?
- Will eventually lead us back to discussing things in terms of “mental spaces”
Language: Compositional Semantics

- What parts of language affect meaning.
- Word order
- Hierarchical structure lets us “make infinite use of finite means” - Chomsky

\[ X \rightarrow \text{this} \]
\[ Y \rightarrow \text{marmot} \]
\[ S \rightarrow XY \]
\[ Z \rightarrow \text{big} \]
\[ Z \rightarrow \text{brown} \]
\[ X*Y \rightarrow XZY \]
Language: Compositional Semantics

Figure 7.4 — *Left:* a tiny grammar, written in the form of production rules. The grammar consists of symbols for variables (S, sentence seed; NP, noun phrase; VP, verb phrase; Det, determiner, etc.) and terminals (Trotzky, the, hyrax, etc.), and rules that allow certain strings to be substituted for others (note that the second rule is recursive). Symbols following each other stand for concatenation; the symbol | denotes disjunction. Repeated application of such rules causes strings of terminals (sentences) to be generated. *Right:* one of the sentences that can be generated by this grammar, and its hierarchical tree structure, which stems from successive invocation of the rules.
Language Learning: Where do we start?

Going digital

- Converting the analog speech stream into discrete elements (phonemes, words, sentences)
- Learning must be unsupervised (there is not sufficient feedback about all these individuals to be “taught” these patterns)
- Can rely on similar computational principals as we have already studied in perception!!
  - Unitization
  - Segmentation
  - Categorical Perception
  - Dimensionality reduction

Going recursive

- Alignment and comparison
- Means putting into correspondence similar functional or perceptual parts of different percepts and comparing what is common or different across them
- Collocation (syntagmatic relationships) e.g. the Monty Python example of “What’s that then?” (see also variational sets)
- Distributional equivalence (words that are similar due to appearing in the same contexts).
Language Learning: Where do we start?

Collocation

What’s all this, then?
What’s all this, then, amen?
What’s all this, then? Mmh. etc...

Distributional Equivalence

I saw the dog chase the cat.
I heard the cat chase the dog.
I felt the cat chase the dog.

I stopped the dog from chasing the cat.
I licked the cat.
I teased the cat.

What’s all this, then?
Figure 7.9 — The unsupervised algorithm of Solan, Horn, Ruppen, and Edelman (2005) distills a grammar from raw text by detecting context-specific, statistically significant collocations and distributional regularities, adding the newly acquired structures to the lexicon, and iterating until all significant regularities have been detected. Its operation is illustrated here on examples from the Air Traffic Information System (ATIS) corpus (Hemphill et al., 1990). Left top: a collocation round trip is detected and assigned a unique pattern code, P1156. Left middle: three words are found to be distributionally equivalent in the context of P1179; they become members in a new equivalence class E1180. Left bottom: the process is iterated with P1156 and E1180 as new full-fledged members of the lexicon, leading to the discovery of a new pattern, P1342. Right: the build-up of hierarchically structured patterns by iterative application of the algorithm. Such patterns can be cast in the form of a grammar (Table 7.3 on the following page). For details of the algorithm, and a discussion of the relationship between the grammars it learns and various linguistic theories, see (Edelman, 2007; Solan et al., 2005).
In the grammar-based view of language, ungrammatical constructions are impossible, while grammatical constructions are accepted in an unquestioning way.

In contrast, language appears more graded. Edleman uses the example of awkwardly worded but understandable spam email to show that grammaticality seems to be one of degree.

This is accommodated in terms of a probabilistic grammar.
The probabilities can be learned from experience!

Figure 7.4 — *Left*: a tiny grammar, written in the form of production rules. The grammar consists of symbols for variables (S, sentence seed; NP, noun phrase; VP, verb phrase; Det, determiner, etc.) and terminals (Trotsky, the, hyrax, etc.), and rules that allow certain strings to be substituted for others (note that the second rule is recursive). Symbols following each other stand for concatenation; the symbol | denotes disjunction. Repeated application of such rules causes strings of terminals (sentences) to be generated. *Right*: one of the sentences that can be generated by this grammar, and its hierarchical tree structure, which stems from successive invocation of the rules.
Language: The Importance of Structure

### n-Gram models

| a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
| a | 60 | 22 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| b | 46 | 0 | 7 | 3 | 2 | 5 | 2 | 2 | 1 | 1 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 1 | 3 | 0 | 1 |
| c | 1 | 59 | 0 | 1 | 1 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 0 |
| d | 3 | 4 | 54 | 8 | 2 | 4 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| e | 30 | 1 | 3 | 67 | 1 | 1 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 1 | 1 | 1 | 0 |
| f | 38 | 12 | 0 | 45 | 5 | 0 | 2 | 1 | 2 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| g | 2 | 2 | 2 | 6 | 1 | 4 | 4 | 4 | 10 | 3 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 0 | 1 | 0 |
| h | 2 | 0 | 0 | 1 | 10 | 3 | 8 | 45 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 0 | 1 | 2 | 9 | 0 | 0 |
| i | 6 | 0 | 5 | 2 | 1 | 6 | 0 | 0 | 61 | 2 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| j | 0 | 2 | 3 | 3 | 1 | 3 | 11 | 2 | 3 | 47 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 3 | 1 | 0 | 1 |
| k | 2 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 1 | 53 | 3 | 10 | 2 | 1 | 0 | 1 | 2 | 0 | 7 | 1 | 0 | 1 | 0 | 1 |
| l | 0 | 1 | 10 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 66 | 0 | 0 | 0 | 9 | 2 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 1 |
| m | 3 | 0 | 8 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 42 | 11 | 0 | 4 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 4 | 0 |
| n | 2 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 9 | 36 | 13 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 0 | 0 | 6 | 4 |
| o | 0 | 1 | 1 | 1 | 4 | 0 | 0 | 2 | 4 | 0 | 2 | 0 | 9 | 36 | 13 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 6 | 4 |
| p | 0 | 1 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 7 | 1 | 1 | 2 | 4 | 4 | 5 | 0 | 3 | 9 | 0 | 1 | 0 | 0 | 1 | 1 |
| q | 3 | 1 | 2 | 1 | 0 | 2 | 3 | 0 | 0 | 1 | 2 | 6 | 0 | 1 | 0 | 9 | 38 | 10 | 1 | 2 | 0 | 0 | 1 | 2 | 3 | 2 |
| r | 1 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 9 | 0 | 1 | 2 | 2 | 1 | 2 | 56 | 2 | 1 | 0 | 4 | 0 | 0 | 2 | 2 |
| s | 2 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 5 | 0 | 0 | 0 | 65 | 4 | 1 | 0 | 0 | 0 | 1 | 0 |
| t | 0 | 3 | 0 | 0 | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 3 | 5 | 1 | 12 | 52 | 0 | 1 | 0 | 0 | 0 | 0 | 4 |
| u | 4 | 0 | 1 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 63 | 2 | 0 | 3 | 2 | 2 |
| v | 1 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 7 | 2 | 1 | 0 | 1 | 0 | 4 | 1 | 0 | 1 | 63 | 0 | 0 | 1 | 2 |
| w | 2 | 3 | 0 | 4 | 0 | 1 | 1 | 0 | 4 | 1 | 8 | 0 | 2 | 2 | 2 | 0 | 3 | 2 | 2 | 0 | 39 | 0 | 6 | 1 | 6 |
| x | 3 | 1 | 4 | 1 | 2 | 2 | 1 | 1 | 1 | 3 | 1 | 6 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 1 | 0 | 39 | 0 | 2 |
| y | 1 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 4 | 1 | 2 | 3 | 3 | 9 | 0 | 0 | 0 | 3 | 0 | 5 | 1 | 0 | 40 | 6 |
| z | 0 | 1 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 1 | 2 | 0 | 0 | 6 | 12 | 0 | 1 | 5 | 0 | 0 | 6 | 3 | 1 | 0 | 4 | 44 |

### Hidden Markov Models/ADIOS/SRNs/etc

- M1
- M2
- M3
- M4
- M5
- M6
- M7
- M8
- M9
- M10

### Flexible search for structure

Primitives are constructed not given

Probabilistic and graded
Language: The Importance of Context

Watch this video!

David E. Rumelhart Prize talk, August 3, 2007
Language: The Neurocomputational Level

• Only humans have language

• Thus, there has long been a belief that language is in some sense special as a natural phenomena

• Advocates of this view resist the idea that we can think of language in the same way that we talk about other behavioral/neuroscientific phenomena like memory, learning, categorization, and perception

• Edelman goes so far to call it equivalent to proponents of “intelligent design” theory
It is perfectly safe to attribute this development [of innate language structures] to "natural selection", so long as we realize that there is no substance to this assertion, that it amounts to nothing more than a belief that there is some naturalistic explanation for these phenomena. [Noam Chomsky, Language and Mind, 1972, p. 97]

In studying the evolution of mind, we cannot guess to what extent there are physically possible alternatives to, say, transformational generative grammar, for an organism meeting certain other physical conditions characteristic of humans. Conceivably, there are none -- or very few -- in which case talk about the evolution of the language capacity is beside the point. [Chomsky 1972 p. 98]

It surely cannot be assumed that every trait is specifically selected. In the case of such systems as language or wings it is not even easy to imagine a course of selection that might have given rise to them. A rudimentary wing, for example, is not "useful" for motion but is more of an impediment. [Noam Chomsky Language and Problems of Knowledge: the Managua Lectures 1988 p 167]

It may be that at some remote period a mutation took place that gave rise to the property of discrete infinity, perhaps for reasons that have to do with the biology of cells, to be explained in terms of properties of physical mechanisms, now unknown. . . . Quite possibly other aspects of its evolutionary development again reflect the operation of physical laws applying to a brain of a certain degree of complexity. [Chomsky 1988, p. 170]
Learning language

Is it possible?
Learning language

- A classic battle-ground of theories of learning (indeed Chomsky’s review of B.F. Skinner’s *Verbal Behavior* book heralded the “end” of the behaviorist paradigm and the start of the “cognitive revolution”)

- Let’s start simple... *language learning happens*:
  - We speak different languages, and it depends on what country or household you were raised in
  - Spanish babies can learn to speak English, English babies can learn to speak Japanese, so obviously there is some role for learning in the acquisition of language!
...however, “poverty of the stimulus”

- Children make grammatical generalization that go beyond what is justified by the evidence in the input (Chomsky, 1965, 1980)
- The evidence in the input is not enough to rule out particular sentences
...however, “poverty of the stimulus”

(1a) The man was hungry.
(1b) Was the man hungry?

(2a) The boy is smiling.
(2b) Is the boy smiling?

(3a) The little girl in the red dress is smiling.
(3b) Is the little girl in the red dress smiling?

(4a) The boy who is smiling is happy.
*(4b) Is the boy who smiling is happy?
(4c) Is the boy who is smiling happy?
...however, “poverty of the stimulus”

(5) Interrogatives can be formed by moving the leftmost auxiliary in the declarative to the beginning of the sentence.

(6) Interrogatives can be formed by moving the auxiliary in the main clause of the declarative to the beginning of the sentence.
...however, “poverty of the stimulus”

- The previous example demonstrates that language representation may contain some hierarchical structure (phase structures).

- Indeed, some argue that human language is capable of “infinite recursion” (i.e., Context-free grammars).

- Gold (1967) proved that it is impossible for a certain, but general class of learning algorithms to acquire context-free grammars on the basis of positive evidence alone.
What is positive and negative evidence?

- **Positive evidence** is examples of a sentence spoken by other members of a language community.

- **Negative evidence** would be feedback indicating which sentences are ungrammatical or unlicensed by the language.

- Numerous arguments have been put forward that negative evidence is sparse at best, and learning a language must go primarily based on positive evidence.
Negative evidence in language acquisition*

Gary F. Marcus

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received February 7, 1992, final version accepted July 2, 1992

---

**Observed proportion of reply instances out of child utterances of a given sentence**

---

Likelihood of obtaining $p(\text{obs})$ given $n = 10$

- $Pr(\text{gr}) = .12$
- $Pr(\text{ungr}) = .20$

---

Likelihood of obtaining $p(\text{obs})$ given $n = 446$

- $Pr(\text{gr}) = .12$
- $Pr(\text{ungr}) = .20$

---

Observed proportion of reply instances out of child utterances of a given sentence
Negative evidence in language acquisition*

Gary F. Marcus
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received February 7, 1992, final version accepted July 2, 1992

Minimum number of times \( n \) a child would need to repeat a given sentence verbatim to decide whether it was grammatical, with chance of error, \( e \), less than .01, calculated for the strongest reply types reported in the discourse studies (\( n \) may be determined by finding the minimum number of repetitions of a single sentence such that the 99\% confidence intervals of the two distributions \( pr\{gr\}, pr\{ungr\} \) do not overlap)

<table>
<thead>
<tr>
<th>Study</th>
<th>Reply type</th>
<th>( pr{ungr} )</th>
<th>( pr{gr} )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hirsh-Pasek et al.</td>
<td>Repetitions 2-year-olds</td>
<td>20</td>
<td>12</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>3–5-year-olds</td>
<td>no differences</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>Penner</td>
<td>Expansions: group 1</td>
<td>18.3</td>
<td>4.6</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>Expansions: group 2</td>
<td>11.3</td>
<td>6.3</td>
<td>679</td>
</tr>
<tr>
<td>Bohannon and Stanowicz</td>
<td>Total of all reply types</td>
<td>35</td>
<td>14</td>
<td>85</td>
</tr>
<tr>
<td>Morgan and Travis</td>
<td>Expansions (A, E, and S)</td>
<td>11.3</td>
<td>2.7</td>
<td>169</td>
</tr>
</tbody>
</table>

Note: The data from Demetras et al. are excluded because most patterns of reply varied across parents and no inferential statistics (or sample sizes) were provided. Without such information we cannot infer the reliability of the data. The data for Morgan and Travis are an average across the patterns of parental expansions to Adam, Eve, and Sarah, from Table 4, Morgan and Travis (1989, p. 545), because there is substantial variation between the pattern of expansions to the different children. Eve would only require 63 verbatim repetitions, but Sarah would require 300.
Marcus concludes that negative feedback is too inconsistent and sparse to be useful

It doesn’t apply equally for all children

The nature of the feedback varies considerably

It is noisy and unpredictable

Thus, **sentence-by-sentence** acceptance tests are unlikely to be an efficient route to learning language
An alternative approach

- Probabilistic inference of grammatical structures!
Next time

**Social learning:** imitation, copying behavior, mirror neurons, etc..