It’s the exceptions that prove the rule or when & why Adam first “feeled dat”

6.S077 Recitation 8 (Week 9)
The drive to learn rules

Berko, 1957: “Wug” test

How much do children “over-regularize”?
This is a KAZH.

Now there is another one.
There are two of them.
There are two _____. 
This is a GUTCH.

Now there is another one.
There are two of them.
There are two ______.
Rules are inevitable due to the infinity of language and the sparsity of examples.
But rules don’t always rule… Sometimes there are exceptions
Why? When? How to generalize?

Only 1 child in 86 said “Glung”
The other children were quiet....

This is a man who knows how to Glung. He is Glunging. He did the same thing yesterday. What did he do yesterday? Yesterday he ________.
gentif mawg
gentif lepal
gentif mawg ka
gentif tomber ka
gentif tomber
gentif lepal
gentif lepal po
We have already seen:
2 regular examples
1 “exception”

Will a child generalize or not?
What about you?
When do kids form rules? Allow exceptions?

- Must allow exceptions! “All grammars leak”
- Example:
  
  Do you have any wool?
  Don’t ask what you can ...

- English since about 1680: use do (and don’t put the verb in front)
- But, there are some common examples (even in child input) that still put the verb at the front):
  
  Baa, baa, black sheep have you any wool?
  Ask not what you can do for your country ...

- But how many exceptions are allowed before we give up and memorize everything as a list?
Case study: regular/vs irregular verbs & past tense

• English regular past tense for verbs: add “d”:
  - bake-baked; cook-cooked; feel-felt; ...; walk-walked; zoom-zoomed (google-googled)

• About 150 exceptions:
  - dig-dug; eat-ate; ring-rang; sing-sung (“irregular” verbs)

• What do kids do?

• Classic pattern
Figure 4.1
Schematic illustration of Adam's U-shaped past-tense development, adapted from Marcus et al. 1992. All irregular verbs were used perfectly until 2;11, when the first instance of overregularization took place.

His vocabulary even though the past-tense form walked may not have made an appearance. This method produced a list of 300 verbs, of which 57 are irregular. In order for the -d suffix to be productive, there should be more than 300 or 53 irregulars. We are agonizingly close (53 vs. 57), and the difference can well be attributed to sampling effects. Since the CHILDES transcripts could not have recorded Adam's complete vocabulary, the regular verbs, which are in general less frequent than the irregulars, must have been undersampled to a greater extent. The more general point of this exercise is to highlight the critical condition for productivity. Only an overwhelming majority will suffice, and Adam rectified the "-d" rule only after the regulars thoroughly outnumbered the irregulars.

The quantitative considerations of the Tolerance Principle provide a concrete and rigorous approach to the problem of individual variation in language acquisition. That individuals in a linguistic community generally learn comparable grammars would follow if the composition of their vocabularies at the conclusion of acquisition is also comparable. That is, the N's and e's across Adam’s "What dat feel like" age 2;11 Why then? Why not before?

"The U-shape pattern"
Possible error types for forming past tense of verbs

• go-gooled (extend regular rule to an irregular verb → over-regularization)
• bring-brang (extend irregular subgroup to irregular verb, compare single-sang → over-irregularization)
• (extend irregular to some entirely new form)
• Remember Sussman-Yip? How is the past tense rule formed?
What about learning the English past tense rule? Sussman & Yip

Child hears “...walk...”; ... “walked”

Rule #1
for: walk
do: -d

Rule #2
for: talk
do: -d

Rule #3
for: *alk
do: -d

Generalize by intersecting (conservatively – why conservative?)

Child hears “...talk...”; ... “talked”
Upon the presentation of English words in the stem and preterite form, the learning model, then, the generality of a rule is directly related to the diversity of words it applies to. This may account for the purported default rule: anything goes. In the present discussion, and also a logical possibility as well as an empirical verification of words it applies to.

As more items are presented incrementally, the condition for the application of the regular rule ("-ed") becomes broader. Eventually, the learner concludes that the "irregular" rule. Once its productivity is established, the "irregular" rule. Once its productivity is established, the "irregular" rule applies everywhere as in Rule 11 in Figure 3.1), for every verb to which it apply (Alegre and Gordon 1999; Baayen et al. 2003), although see Lignos and Gorman 2012 for a reinterpretation.

Default rule for past tense – applies everywhere (most general) – might be too general.
Regular rule plus a small set of special case “subrules” for some irregular verbs

This actually uses “distinctive features” – the “add d” rule is really this:

→ d / [+voice, −coronal] (close-clos/d/)
→ t / [−voice, +strident] (bake-bake/t/)

1. Internal syllable → ɔt/ (e.g., think-thought, catch-caught, buy-bought)
2. i → æ / N (sing-sang, ring-rang, ...)
3. i → e / C [d, t] (feed-fed, lead-led, meet-met, ...)
4. [aI, O, o] → u/ [l, n, r] (fly-flew, grow-grew, know-knew, draw-drew, ...)
5. Regular rule: Otherwise: apply the d/t rule

But this is not perfect - Where does this kind of learner go wrong?
Over- and under-generalization for Sussman-Yip model

• Overgeneralizes: “special” rule 1 is too general; would convert “love” to “lought” (why else is overgeneral rule possibly bad?)

• Yip-Sussman solves only half the learning problem: Does OK at figuring out the form of rule, but not so good in figuring out when the rules should apply (the scope of rules)
“Elsewhere” principle for rule systems

• How to order the rules? Order such that the most general (least specific) rule is last
• Why?
• Suppose you put most specific rules at the end...what would happen?
• How to order this list of exceptions?
• Order by frequency – keep ordered list of exceptions ranked by frequency, most frequent first
• Not hard to do: “move up” algorithm (Rivest, 1976) – swap just-used item with one just above it (cf. smart phone app list)
Upshot: Kids like it fast

The “tipping point”: suppose we want fastest match

Rules + e exceptions

Exception 1
Exception 2
...
Exception e
Rule \((N-e) \text{ times}\)

<

Just exceptions

Exception 1
Exception 2
...
Exception \(N\)

Run
Time \(T(N, e)\)

Run
Time \(T(N, N)\)

Time \(T\) to process rule if \(N\) items with \(e\) exceptions = \(T(N, e)\)

Time \(T\) to process list if \(N\) items with \(N\) exceptions = \(T(N, N)\)

Which is faster?

when is \(Time(N, e) \leq Time(N, N)\)? \(\text{“Tipping Point”}\)
The Tipping Point: when to use a rule rather than all special case memorized exceptions, given $N$ examples

If $Time(N, e) \leq Time(N, N) \rightarrow$ **Using Rule is faster** than listing all examples (a *productive* rule)

The *equality* holds at the *tipping point*: $N/\ln N$

**SO:**

- if we are *below* this tipping point, having a rule is faster than just memorizing;
- if we are *above* this point, memorizing full list is faster.

Example: $N = 9$, $N/\ln N = 9/2.19 = 4.096$, so if fewer than 4 exceptions $\rightarrow$ use rule;
if 4 or more exceptions $\rightarrow$ memorize
Expected search time

\[ e \leq \theta_N := \frac{N}{\ln N} \]

Flat green line = time if all \( N \) examples memorized as a list (all exceptions)
Red line = time if a rule with \( e \) exceptions

Rule is faster if \( \leq 22 \) exceptions
\((100/(\ln 100) = 100/4.6 = 21.78)\)
Tipping point for # of Exceptions $\theta$ given some # of items $N$, $\theta_N=(N/\ln N)$

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\theta_N$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>40.0</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>35.0</td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>26.0</td>
</tr>
<tr>
<td>100</td>
<td>23</td>
<td>23.0</td>
</tr>
<tr>
<td>200</td>
<td>38</td>
<td>19.0</td>
</tr>
<tr>
<td>500</td>
<td>80</td>
<td>16.0</td>
</tr>
<tr>
<td>1,000</td>
<td>145</td>
<td>14.5</td>
</tr>
<tr>
<td>5,000</td>
<td>587</td>
<td>11.7</td>
</tr>
</tbody>
</table>

Not majority rules!
Smaller # of items – greater fraction of exceptions tolerated before rule judged unproductive, and cheaper to just list all the items – Small data = more rules

And this explains the appearance of rules, and their over-application, etc. in the child language database
Most frequent verbs (total #, # irregular exceptions)

bold font = point at which # irregular < tipping pt θ

<table>
<thead>
<tr>
<th>Top N</th>
<th>sing→sang</th>
<th>feed→fed</th>
<th>fly→flew</th>
<th>-d</th>
<th>θ_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
<td></td>
<td>(8, 3)</td>
<td>(100, 54)</td>
<td>22</td>
</tr>
<tr>
<td>200</td>
<td>(3, 1)</td>
<td></td>
<td>(11, 5)</td>
<td>(200, 76)</td>
<td>37</td>
</tr>
<tr>
<td>300</td>
<td>(3, 1)</td>
<td></td>
<td>(13, 8)</td>
<td>(300, 92)</td>
<td>53</td>
</tr>
<tr>
<td>500</td>
<td>(5, 2)</td>
<td>(6, 3)</td>
<td>(15, 10)</td>
<td>(500, 103)</td>
<td>80</td>
</tr>
<tr>
<td>800</td>
<td>(8, 5)</td>
<td>(11, 7)</td>
<td>(18, 13)</td>
<td>(800, 121)</td>
<td>119</td>
</tr>
<tr>
<td>1022</td>
<td>(8, 5)</td>
<td>(13, 9)</td>
<td>(22, 16)</td>
<td>(1022, 127)</td>
<td>147</td>
</tr>
</tbody>
</table>

In first 100 most frequent verbs, 54 have irregular past tense, e.g., go-went, so no chance of rule

But 8 fly-flew type verbs, with 5 examples and 3 exceptions < than t.p. 8/ln 8 = 3.8 → form a rule

*blow-blew, cry-cried, fly-flew, grow-grew, know-knew, lie-lied, throw-throw, try-tried*
Most frequent verbs (total #, # irregular exceptions)
bold font = point at which # irregulars < tipping pt θ

<table>
<thead>
<tr>
<th>Top N</th>
<th>sing—sang</th>
<th>feed→fed</th>
<th>fly→flew</th>
<th>-d</th>
<th>θ_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>—</td>
<td>—</td>
<td>(8, 3)</td>
<td>(100, 54)</td>
<td>22</td>
</tr>
<tr>
<td>200</td>
<td>(3, 1)</td>
<td>—</td>
<td>(11,5)</td>
<td>(200, 76)</td>
<td>37</td>
</tr>
<tr>
<td>300</td>
<td>(3, 1)</td>
<td>—</td>
<td>(13, 8)</td>
<td>(300, 92)</td>
<td>53</td>
</tr>
<tr>
<td>500</td>
<td>(5, 2)</td>
<td>(6, 3)</td>
<td>(15, 10)</td>
<td>(500, 103)</td>
<td>80</td>
</tr>
<tr>
<td>800</td>
<td>(8, 5)</td>
<td>(11, 7)</td>
<td>(18, 13)</td>
<td>(800, 121)</td>
<td>119</td>
</tr>
<tr>
<td>1022</td>
<td>(8, 5)</td>
<td>(13, 9)</td>
<td>(22, 16)</td>
<td>(1022, 127)</td>
<td>147</td>
</tr>
</tbody>
</table>

Top 200 verbs: Now 11 fly-flew verbs, with 5 exceptions and θ₁₁ = 4, so # exceptions > tipping point 4, so **no rule**

*blow-blew, cry, fly-flew, follow, grow-grew, know-knew, lie, snow, swallow, throw-threw, try*

This generalization would disappear, but...

There are now 3 sing-sang, verbs, with 1 exception, so # exceptions < θ₃ = 2.7

*Bring-brought, ring-rang, sing-sung*, so the sing-sang rule can persist for a bit
Why Adam “feeled” at age 2 years, 11 months? “What dat feeled like”

• Entire database: 1022 unique verbs in past tense, 127 are irregular
• $\theta_{1027} = 147$ (so room for more irregulars)

Productivity as function of vocabulary size – top $N$ past tense verbs in child corpus
- $d$ column: (# regular verbs, # exceptions found); $\theta_N$ = # exceptions tolerated

<table>
<thead>
<tr>
<th>Top $N$</th>
<th>sing→sang</th>
<th>feed→fed</th>
<th>fly→flew</th>
<th>$-d$</th>
<th>$\theta_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>—</td>
<td>—</td>
<td>(8, 3)</td>
<td>(100, 54)</td>
<td>22</td>
</tr>
<tr>
<td>200</td>
<td>(3, 1)</td>
<td>—</td>
<td>(10, 5)</td>
<td>(200, 76)</td>
<td>37</td>
</tr>
<tr>
<td>300</td>
<td>(3, 1)</td>
<td>—</td>
<td>(13, 8)</td>
<td>(300, 92)</td>
<td>53</td>
</tr>
<tr>
<td>500</td>
<td>(5, 2)</td>
<td>(6, 3)</td>
<td>(15, 10)</td>
<td>(500, 103)</td>
<td>80</td>
</tr>
<tr>
<td>800</td>
<td>(8, 5)</td>
<td>(11, 7)</td>
<td>(18, 13)</td>
<td>(800, 121)</td>
<td>119</td>
</tr>
<tr>
<td>1022</td>
<td>(8, 5)</td>
<td>(13, 9)</td>
<td>(22, 16)</td>
<td>(1022, 127)</td>
<td>147</td>
</tr>
</tbody>
</table>

As long as $\theta_N < # \text{ actual exceptions}$, regular rule is not accepted

Example: *Fly-flew*
1. *blow*, cry, *fly*, *grow*, *know*, lie, *throw*, try ($\theta_8$=3)
Figure 4.1
Schematic illustration of Adam’s U-shaped past-tense development, adapted from Marcus et al. 1992. All irregular verbs were used perfectly until 2;11, when the first instance of overregularization took place.

His vocabulary even though the past-tense form walked may not have made an appearance. This method produced a list of 300 verbs, of which 57 are irregular. In order for the -d suffix to be productive, there should be more than 300 or 53 irregulars. We are agonizingly close (53 vs. 57), and the difference can well be attributed to sampling effects. Since the CHILDES transcripts could not have recorded Adam’s complete vocabulary, the regular verbs, which are in general less frequent than the irregulars, must have been undersampled to a greater extent. The more general point of this exercise is to highlight the critical condition for productivity. Only an overwhelming majority will suffice, and Adam rectified the “-d” rule only after the regulars thoroughly outnum-bered the irregulars.

The quantitative considerations of the Tolerance Principle provide a concrete and rigorous approach to the problem of individual variation in language acquisition. That individuals in a linguistic community generally learn comparable grammars would follow if the composition of their vocabularies at the conclusion of acquisition is also comparable. That is, the N’s and e’s across Adam “What dat feeled like” age 2;11 Why then? Why not before?

The U-shape pattern

Adam

% Irregular Past Tense Correct

0 10 20 30 40 50 60 70 80 90 100

2;0 2;2 2;4 2;6 2;8 2;10 3;0 3;2 3;4 3;6 3;8 4;0 4;2 4;4 4;6

Age

“What dat feeled like” age 2;11 Why then? Why not before?
Why Adam feeled at age 2;11 but not before

### Table 4.1

<table>
<thead>
<tr>
<th>Top N</th>
<th>sing→sang</th>
<th>feed→fed</th>
<th>fly→flew</th>
<th>-d</th>
<th>$\theta_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>—</td>
<td>—</td>
<td>(8, 3)</td>
<td>(100, 54)</td>
<td>22</td>
</tr>
<tr>
<td>200</td>
<td>(3, 1)</td>
<td>—</td>
<td>(10, 5)</td>
<td>(200, 76)</td>
<td>37</td>
</tr>
<tr>
<td>300</td>
<td>(3, 1)</td>
<td>—</td>
<td>(13, 8)</td>
<td>(300, 92)</td>
<td>53</td>
</tr>
<tr>
<td>500</td>
<td>(5, 2)</td>
<td>(6, 3)</td>
<td>(15, 10)</td>
<td>(500, 103)</td>
<td>80</td>
</tr>
<tr>
<td>800</td>
<td>(8, 5)</td>
<td>(11, 7)</td>
<td>(18, 13)</td>
<td>(800, 121)</td>
<td>119</td>
</tr>
<tr>
<td>1022</td>
<td>(8, 5)</td>
<td>(13, 9)</td>
<td>(22, 16)</td>
<td>(1022, 127)</td>
<td>147</td>
</tr>
</tbody>
</table>

So this is almost exactly when we would expect the regular productive rule to start being applied, and it marks the first appearance of over-application of the rule in Adam’s speech.

Look at Adam’s use of past tense verbs by the age 2;11 – point of emergence of regular rule.

By this point, Adam had 300 verbs, of which 57 were irregular. $\theta_{300}=53$

So: (nearly) at tipping point for rule!
"Nouns" presented according to a power law Zipf frequency; 27 example sentences; 2 different sequences of regular patterns vs. exceptions

Regular ending: ka
Irregular: po, tay, lee, bae, mau, woo

9 examples, tipping point = 4.078
When will child hold onto rule?

“Nouns” presented according to a power law Zipf frequency; 27 example sentences; 2 different sequences of regular patterns vs. exceptions

16 Children age 6-8 years old
Presented with artificial ‘sentences’ w/ two diff ft mixes of regular vs. exceptions to “plural”

Now test tipping point of 4.078 by creating 2 sets of stimuli, one with regulars > tipping point, one < tipping point: 5, 4 vs. 3, 6
from 16 children age 6-8 years

5 Regular Forms/4 Exceptions (predict: rule)  3 Regular Forms/6 Exceptions (predict: no rule)

\[ \theta_N = \frac{N}{\ln N} = \frac{9}{\ln 9} = 4.096 \]
Results from 16 children age 6-8 years

5 Regular Forms/4 Exceptions

3 Regular Forms/6 Exceptions

5 regular, 4 not; vs. 3 regular, 6 not (total 9 examples)
Why less is more: when Small Data succeeds

Study Reveals: Babies Are Stupid

Above: Despite their relatively large cranial capacities, babies such as this one are so unintelligent that they are unable to distinguish colorful plastic squeak toys from food sources.

clean rules of language

noise!

the long tail