Why 72?

Charles Yang
University of Pennsylvania

NAACL 2018
Our Family-Friendly Chair

Courtesy of Chris Callison-Burch
… because children take off after 72.”
Karen Fuson (p.c., March 2016)
WHAT COUNTS? 18

highest count of 100 (mean = 79.2, SD = 34.29), reflecting the large number of very high counters we have in our sample. Consistent with the predictions of Yang (In prep.), we have not yet found a very high counter (highest count greater than 80) who makes an error in the count list.

Figure 3. Scatterplot relating age to highest count, with participants grouped by count level.

Successor task

In analyzing accuracy on the Successor task, we plan to test participants' accuracy on a given number range against chance (50%) in independent two-tailed t-tests. Thus, we will test performance for low counters on small numbers, and so on. Previous research has found in these tests that low, medium, and high counters generally perform at (Davidson et al., 2012) or slightly above chance (Cheung et al., 2017) for numbers within their count range, and that only very high counters perform significantly above chance for all numbers on which they are tested (Cheung et al., 2017). We seek to replicate this finding in our own sample. Above-chance performance across all number ranges in this task indicates that a child truly understands the Successor Function, and has likely generalized it to numbers outside of her count range. If a child's performance is either at or significantly below chance, Rose Schneider (2016)
The Chinese Advantage

“Four-year-olds in China made very rapid progress in generalizing number names up to 100 after they could count to approximately 40” (Miller, Kelly, & Zhang 2005)
A Roadmap

• How children learn the rules of languages
  • A proposal and an assortment of evidence
• What it can do to help develop unsupervised learning systems
  • Making the most out of very little data
• Why 72
  • Some connection between language and number
Sparsity of Input

Courtesy of Erwin Chan and Constantine Lignos
Richness of Output

From Guasti (1992, Language Acquisition)
A noun ⇐⇒ The noun

hallmark of human language
What children say

The a

Cat car dog boy chair
Statistics of Grammar

A baseline: assume independence and multiply marginal probabilities but take sparsity into account

(Yang 2013, PNAS)
• Typical parental data

• **914** singular nouns, only **34%** are used with both *a* and *the*, but the child used them interchangeably

• A third of the batters are observed to switch-hit

• **All** switch-hitters?
The first three minutes

Bracketing FMeasure, %

% Training Data PTB (sec 23 held out)

$V_{read} P \rightarrow V_{read} NP$

$VP \rightarrow V NP$
Why so slow?

“What that feeled like”?

Pinker (1995, *An invitation to cognitive science*)
Majority doesn’t rule

German noun plural suffixes

-\(n\) 48%
-\(\emptyset\) 17%
-s 4%
-e 27%
-er 4%

Autos, Parks, Pizzas, …, iPhones

Clahsen (1999, *Brain and Behavioral Sciences*)

English word stress

Noninitial 16%
Initial 84%

They permit you to get a permit
They will record a record.

Cutler & Davis (1987, *Comp. & Speech*)
How do we learn rules?
Not so Sure?
• Give a set of items:
  • If many do X, then all do X
  • if few do X, then don’t do X (or withhold judgment)
• How many is many or few?
Rule + Exceptions

- Exception 1
- Exception 2
- Exception 3
- ...
- Exception e
- Rule (N-e)

Memorize Everything

- Exception 1
- Exception 2
- Exception 3
- ...
- ...
- Exception N

\[ \sum_{i=1}^{e} ip_i + iP[N - e] \]

\[ \sum_{i=1}^{N} ip_i \]
Tolerance Principle

A productive rule over \( N \) items cannot have more than \( \frac{N}{\ln N} \) exceptions.
Tolerance Principle in Action

<table>
<thead>
<tr>
<th>N</th>
<th>$\theta_N$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>40%</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>35%</td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>26%</td>
</tr>
<tr>
<td>100</td>
<td>22</td>
<td>22%</td>
</tr>
<tr>
<td>200</td>
<td>38</td>
<td>19%</td>
</tr>
<tr>
<td>500</td>
<td>80</td>
<td>16%</td>
</tr>
<tr>
<td>1000</td>
<td>144</td>
<td>14%</td>
</tr>
<tr>
<td>2000</td>
<td>263</td>
<td>13%</td>
</tr>
<tr>
<td>5000</td>
<td>587</td>
<td>12%</td>
</tr>
<tr>
<td>10000</td>
<td>1086</td>
<td>11%</td>
</tr>
</tbody>
</table>

Parameter free, small is better

Let's start with something very familiar, the acquisition of past tense in English. Suppose the learner has arrived at a set of morphological rules such as those produced by some suitable model of inductive learning (e.g., Yip and Sussman 1997). Those in (1a) are irregular and the learner should eventually assess them to be unproductive while the rule in (1b) is regular and the learner should recognize it as such.

1 When the teacher holded the baby rabbits

2 For simplicity of presentation, we omit the phonological alternation for the -d suffixation, which can be automatically induced by the learning model as shown in Chapter 3.
The "wug" test revisited (adapted from Berko, 1958)

This is a man who knows how to GLING. He is GLINGING. He did the same thing yesterday. What did he do yesterday? Yesterday he _______.

Percentage Children

child performance

<table>
<thead>
<tr>
<th></th>
<th>wugs</th>
<th>bik's</th>
<th>zibbing</th>
<th>ricked</th>
<th>glang</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

This is a WUG
Now there is another one. There are two of them. There are two ______.
Artificial Language

Why 5/4 and 3/6?

9/ln9 = 4.2!

Conditions

<table>
<thead>
<tr>
<th>Conditions</th>
<th>5R4E</th>
<th>3R6E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ka</td>
<td>ka</td>
</tr>
<tr>
<td></td>
<td>ka</td>
<td>ka</td>
</tr>
<tr>
<td></td>
<td>ka</td>
<td>ka</td>
</tr>
<tr>
<td></td>
<td>ka</td>
<td>po</td>
</tr>
<tr>
<td></td>
<td>lee</td>
<td>bae</td>
</tr>
<tr>
<td></td>
<td>bae</td>
<td>tay</td>
</tr>
<tr>
<td></td>
<td>muy</td>
<td>woo</td>
</tr>
</tbody>
</table>

KA: regular
E: irregulars

Training

The experimental paradigm

Day 1 (Schuler, Yang, & Newport, 2016a)
Testing

Experimenter says "gentif norg."

Child says "gentif ____"
Essentially categorical!

15 children age 6-8 years
Making it harder

**Conditions**

<table>
<thead>
<tr>
<th></th>
<th>5R4E</th>
<th>3R6E</th>
</tr>
</thead>
<tbody>
<tr>
<td>ka</td>
<td>po</td>
<td>ka</td>
</tr>
<tr>
<td>ka</td>
<td>lee</td>
<td>ka</td>
</tr>
<tr>
<td>ka</td>
<td>bae</td>
<td>ka</td>
</tr>
<tr>
<td>ka</td>
<td>tay</td>
<td>muy</td>
</tr>
<tr>
<td>ka</td>
<td></td>
<td>woo</td>
</tr>
</tbody>
</table>

(Schuler, Yang, & Newport, 2016b)
Broken?

30

(Schuler, Yang, & Newport, 2016b)
Personalized Tolerance Principle

$\frac{8}{\ln(8)} = 3.85$

(Schuler, Yang, & Newport, 2016b)

Experiment 2

Usage of regular form

5R4E  3R6E
Categorical again

Experiment 2

Usage of regular form

Rule

No rule

(Schuler, Yang, & Newport, 2016b)
When *felt* became *feeleed*?

Adam

\[N = 300\]
\[e = 54\]
\[N/\ln N = 52\]

Overregularization

Percent Correct

Vocabulary: Number of Words

1,116 words
(Children in professional families)

749 words
(Children in working class families)

525 words
(Children in welfare families)
Generalization from Small Data

the
a

cat
car
dog
boy
chair

the
a

cat
car
dog
boy
chair

SWITCH HITTER
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.

- All nouns: only 40% appear with both
- Top 50 nouns: 43 appear with both
- Top 100 nouns: 87 appear with both
Unsupervised Morphology

- Mitch Marcus (Penn), Lyle Ungar (Penn)
- Emily Pitler (Google), Erwin Chan (Arizona), Constantine Lignos (ISI), Hongzhi Xu (Penn)

- **er**: productive but with corner cases
  - think-thinker, drink-drinker, …, corn-corner, live-liver

- Young children: “sounder” = radio, “caser” = storage bin

- **et**: spurious and should be purged
  - wall-wallet, bull-bullet, ball-ballet, ass-…
Embedding -er in 4.5 million words

N = 8 + 4 = 12

12 / ln(12) = 5

-er is good

mow — mower
sing — singer
toast — toaster
kick — kicker
think — thinker
hang — hanger
engine — engineer
help — helper
moth — mother
should — shoulder
tim — timer
dang — danger
Embedding -et in 4.5 million words

N = 1 + 6 = 7

-et is bad

rock-rocket
upset—ups
rack—racket
mark—market
blank—blanket
ball—ballet

cabin—cabinet
Approach

- Morphology = Phonology + Semantics (e.g., Schone & Jurafsky 2001)
- A conventional distributional learner for affixes \{S\}
- Subject each affix (S) to Tolerance test
  - \(N\) is the number of word pairs related by \(S\)
  - Accept \(S\) if there are fewer than \(\frac{N}{\ln N}\) negatively related words
- Segmentation using filtered \(\{S\}'\)
# Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Embedding</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBJ2015</td>
<td>129M Wikipedia</td>
<td>0.807</td>
<td>0.722</td>
<td>0.762</td>
</tr>
<tr>
<td>cccccc-, -cccccc +tolerance</td>
<td>4.5M child-directed English</td>
<td>0.879</td>
<td>0.719</td>
<td>0.786</td>
</tr>
<tr>
<td>Xu et al. 2018</td>
<td>none</td>
<td>0.810</td>
<td>0.787</td>
<td>0.798</td>
</tr>
<tr>
<td>Xu et al. 2018 +tolerance</td>
<td>17M words (Word2Vec demo)</td>
<td>0.840</td>
<td>0.770</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Narasimhan, Barzilay, & Jaakkola (2015, *TACL*)

Xu, Marcus, Ungar, & Yang (2018, *COLING*)

Thanks to Hongzhi Xu
Why 72?

one two three four five six seven eight nine ten

eleven twelve thirteen fourteen fifteen sixteen seventeen

eighteen nineteen twenty 21 22 23 24 25 26 27 28 29

thirty 31 32 33 34 35 36 37 38 39

40 41 42 43 44 45 46 47 48 fifty

51 52 53 54 55 56 57 58 59 60 …

\[
\frac{73}{\ln 73} = 17
\]
Why approximately 40?

“Four-year-olds in China made very rapid progress in generalizing number names up to 100 after they could count to approximately 40” (Miller, Kelly, & Zhang 2005)
Are there 8 or 9 fish in Mr. Dino’s bucket?

The task was presented on a tablet. An example of the stimuli presentation can be seen in Figure 1. At the beginning of each trial, the experimenter introduced a studied animal (“Mr. Dino”) who liked to put fish in the bucket on the screen. The experimenter explained that sometimes Mr. Dino had trouble figuring out how many fish were in the bucket. Once the experimenter was sure the child understood the premise, she advanced the experiment, displaying a number of fish. In the training trial, the experimenter said “Look! I have three fish! I’m putting three fish in Mr. Dino’s bucket,” and advanced the experiment to show the three fish jumping into the bucket so that children did not have visual access to the number of fish. After the fish were occluded the experimenter performed a memory check, asking “How many fish are in the bucket?” If the child did not successfully answer the memory check, the experimenter said “Whoops, let’s try that again!” and started the trial from the beginning. This process was repeated until the child successfully completed the memory check.

Next, the experimenter said “Look!”, directing the child’s attention back to the screen. The experimenter advanced the trial, causing either one or two fish to appear on the screen, and then jump into the bucket. After the fish were no longer visible, the experimenter asked “Are there $N+1$ or $N+2$ fish in the bucket now?” Thus, in the training trial, the experimenter asked “Are there four or five fish in the bucket now?”

The maximum number of trials was 12. The order of the trials was pseudo-randomized by count level, such that the first trials children saw were within their count range.
Productive Counting is the Key

40-79

80-100

(c) High Counters (n = 19)

(d) Very High Counters (n = 34)

23, 24, 28, 31, 35, 36

53, 57, 76, 77

Cheung et al. (2017, Cog. Psych.)
The Induction of Successor Function

- **Base**: $1+1=2$ (infants and other animals know this innately)

- **Induction**: what’s true of the first two elements is true of all elements in an infinitely ordered list
  - Can only do so by counting sufficiently high, in order to work out the productive **rule** of counting

- The concept of integers and discrete infinity is enabled by our linguistic ability to form rules
  - Only compelling case for Whorf (Pica et al. 2004, *Science*)
“the Munduruku do not have a counting routine ... By requiring an exact one-to-one pairing of objects with the sequence of numerals, counting may promote a conceptual integration of approximate number representations, discrete object representations, and the verbal code” (Pica et al. 2004)
Conclusion

• Children can find language from very small data so should we
  • Big data may even be harmful

• The cognitive processes of language acquisition may be deductively studied and contribute to the development of unsupervised NLP systems