A Statistical Test for Grammar

Charles Yang
Department of Linguistics & Computer Science
Institute for Research in Cognitive Science
University of Pennsylvania

CMCL 2011
Portland, OR
Saying and Knowing

• What one says might know reflect what knows about language
  • Chomsky (1965), C. Chomsky (1969, R. Brown (1973) on Adam, Eve, and Sarah against Braine (1963)’s Pivot Grammar
  • Also the early work of, Bellugi, Bloom, Bowerman, Cazden, Fraser, McNeill, Schlesinger, Slobin among others
  • Shipley, Smith & Gleitman (1969, *Language*) on telegraphic speech
  • Not everything that one knows will be said (e.g., islands)
  • Competence/performance
The usage-based turn

• “(w)hen young children have something they want to say, they sometimes have a set expression readily available and so they simply retrieve that expression from their stored linguistic experience” (Tomasello 2000, 77)

• Chief evidence: limited range of combinatorial diversity

• **Verb Island Hypothesis** (Tomasello 1992): “Of the 162 verbs and predicate terms used, almost half were used in one and only one construction type, and over two-thirds were used in either one or two construction types.”

• **Morphology** (Pizutto & Caselli 1994): Italian children use only 13% of stems in 4 or more person-number agreement forms.

• **Determiners and nouns**: next slide

• Beginning to influence ACL **but what’s the Null Hypothesis?**
A basic observation

• “give me X”, a highly frequent expression, is often cited as evidence of the child using formulaic expressions

• From the Harvard study (0.5M words)
  • give me: 93, give him: 15, give her: 12, or 7.75 : 1.23 : 1
  • me: 2870, him: 466, her: 364, or 7.88 : 1.28 : 1

• Need to work out a proper baseline
Diversity of Usage: determiner-noun

• Valian (1986): the knowledge of the category **determiner** fully productive by 2;0, virtually no errors

• low error rate could be memory and retrieval

• Pine & Lieven (1997): **overlap** is much lower than, say, even 50% (following Tomasello’s verb island hypothesis)

\[
\text{overlap} = \frac{\# \text{ of nouns with BOTH the AND a}}{\# \text{ of nouns with EITHER the OR a}}
\]

• But Valian, Solt & Stewart (2008, *J. Child Language*) found **no difference** between kids and their mothers!

• Brown corpus: overlap for **the** and **a** is 25.2% < some children in Pine et al.
Zipf’s long tail

\[
\text{rank} = \frac{C}{\text{frequency}} \quad \log(\text{rank}) = \log C - \log(\text{frequency})
\]

- Excellent fit across languages and genres (Baroni 2008)
- Power law like patterns in morphology (Chan 2008), n-grams (Ha et al. 2002), and syntactic rules
- Rules from Treebank; certain functional words are merged
The Grammar Hypothesis

• Assume DP⇒DN is completely **productive**: combination is independent
  
  • D⇒a/the, N⇒cat, book, desk, water, sun ...

  • other phrases/structures can be analyzed similarly

• Given the Zipfian distribution of words, overlap is necessarily low
  
  • Most nouns will be sampled only once in the data: zero overlap

  • If a noun is sampled multiple times, there is still a good chance that it is paired with only **one** determiner, which also results in zero overlap

  • If the determiner frequencies are Zipfian as well, this makes the overlap **even lower**
Imbalanced determiners

- “the bathroom” » “a bathroom”
- “a bath” » “the bath”
- Brown corpus: 75% of singular nouns occur with only the or a
  - 25% of the remainders (6.25% in total) are balanced
  - for those with both, favored vs. less favored = 2.86 : 1
- This is also true of CHILDES data, for both children and adults (12 samples)
  - 22.8% appear with both, favored vs. less favored = 2.54 : 1
- Imbalance is more Zipfian than Zipf (2:1)
Zipfian Probabilities

- The $r$th word has probability of $P_r$

$$\frac{C}{r} \left( \frac{C}{1} + \frac{C}{2} + \ldots + \frac{C}{N} \right)$$

$$\frac{1}{rH_N} \text{ where } H_N = \sum_{i=1}^{N} \frac{1}{i}$$

- We can approximate the occurrences of nouns and determiners in any sample accurately, regardless of their identities
Expected overlap

\[ O(n_r) = 1 - \Pr\{n_r \text{ is not sampled during } S \text{ trials}\} \]

\[ - \sum_{i=1}^{D} \Pr\{n_r \text{ is sampled but with the } i\text{th determiner exclusively}\} \]

\[ = 1 - (1 - p_r)^S \]

\[ - \sum_{i=1}^{D} \left[ (d_i p_r + 1 - p_r)^S - (1 - p_r)^S \right] \]

**Graph**

D=2, N=100, S=200
Empirical Data

• Children: Adam, Eve, Sarah, Nina, Naomi, Peter

• All children in CHILDES that started at one/two word stage and with reasonably large longitudinal samples

• Used a variant of the Brill tagger (1995) with statistical information for disambiguation: sufficiently adequate due to the unambiguity of “a” and “the”

• Standard procedure in child data processing:
  • remove annotation markers
  • repetitions count only once (“a doggie! a doggie! a doggie!”)
  • extract D-N_{sg} pairs
Empirical and Theoretical Results

also considered the first 100, 300, 500 tokens of the six children (earliest stages of longitudinal development)

paired t- and Wilcoxon tests reveal no difference
Null hypothesis is confirmed

\[ y = 1.08x, \quad R^2 = 0.97 \]

Slight over-estimation due to the Zipf 2:1 ratio for determiners rather than the somewhat more imbalanced empirical ratio.
Why Variation

- Some children have higher overlap than others (and Brown)
  - sample size alone does not predict overlap
- Overlap is determined by how many nouns (out of N) can be expected to be sampled more than once, or
  \[ S \frac{1}{rH_N} > 1 \]

  \[ r = \frac{S}{H_N} \approx \frac{S}{\ln N} \]
- Overlap is a monotonically increasing function of
  \[ \frac{S}{N \ln N} \text{ or } \approx \frac{S}{N} \]
Analysis of Variation

\[ r = 0.986, \ p < 0.00001 \]
Interim Conclusion

- Children’s determiner usage is consistent with the hypothesis of fully productivity from early on.
- This does not tell us how the child arrives at the correct rule.
- This does not mean all aspects of grammar are learned correctly and productively, despite claims in the grammar-based literature.
- We have a statistical test for productivity given limited data.
- It is premature to conclude, based on low overlap data, that child language is item-based.
- Item-based learning needs to make some quantitative predictions about what to expect.
- Extension experiments (e.g., Wug) have severe limitations.
Does memory+retrieval work?

- “... they may appear, and indeed may be, less rigorously specifiable than generative approaches, a disadvantage to some theorists, perhaps.” (Tomasello 1992, p274)

- “(w)hen young children have something they want to say, they sometimes have a set expression readily available and so they simply retrieve that expression from their stored linguistic experience” (Tomasello 2000, 77)

- Tentative approach: model the learner as a list of joint D-N pairs with their associated frequency rather than independently combined units

- **global memory learner**: list consists of 6.5 million words of child-directed speech in the CHILDES database

- **local memory learner**: list consists of the child-directed utterance for each particular child in the CHILDES transcript

- calculate the overlap for the sampled D-N pairs, averaging over 1000 trials
Item-based learners

<table>
<thead>
<tr>
<th>Child</th>
<th>Sample Size (S)</th>
<th>Overlap (BIG learner)</th>
<th>Overlap (small learner)</th>
<th>Overlap (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eve</td>
<td>831</td>
<td>16.0</td>
<td>17.8</td>
<td>21.6</td>
</tr>
<tr>
<td>Naomi</td>
<td>884</td>
<td>16.6</td>
<td>18.9</td>
<td>19.8</td>
</tr>
<tr>
<td>Sarah</td>
<td>2453</td>
<td>24.5</td>
<td>27.0</td>
<td>29.2</td>
</tr>
<tr>
<td>Peter</td>
<td>2873</td>
<td>25.6</td>
<td>28.8</td>
<td>40.4</td>
</tr>
<tr>
<td>Adam</td>
<td>3729</td>
<td>27.5</td>
<td>28.5</td>
<td>32.3</td>
</tr>
<tr>
<td>Nina</td>
<td>4542</td>
<td>28.6</td>
<td>41.1</td>
<td>46.7</td>
</tr>
<tr>
<td>First 100</td>
<td>600</td>
<td>13.7</td>
<td>17.2</td>
<td>21.8</td>
</tr>
<tr>
<td>First 300</td>
<td>1800</td>
<td>22.1</td>
<td>25.6</td>
<td>29.1</td>
</tr>
<tr>
<td>First 500</td>
<td>3000</td>
<td>25.9</td>
<td>30.2</td>
<td>34.2</td>
</tr>
</tbody>
</table>

- Paired t- and Wilcoxon tests show significant differences (p < 0.005)
Measuring Productivity

• The calculation should distinguish productive and unproductive processes as measured by interchangeability.

• An example in morphology

  • \( V_{\text{INFL}} \rightarrow V + \text{suffix} \)

  • suffix \( \rightarrow \text{ed|ing} \)

  • irregulars do not take -ed, so the empirical overlap measure ought to be lower than the theoretical calculation that assumes -ed and -ing are interchangeable (subject to Zipfian frequencies)

• Check the percentage of verbs that take both -ed and -ing.
## -ed vs. -ing

<table>
<thead>
<tr>
<th>File</th>
<th>S</th>
<th># V</th>
<th>Emp.</th>
<th>Theo.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>62807</td>
<td>3044</td>
<td>45.5%</td>
<td>75.6%</td>
</tr>
<tr>
<td>Adam</td>
<td>6774</td>
<td>263</td>
<td>31.3%</td>
<td>90.5%</td>
</tr>
<tr>
<td>Eve</td>
<td>1028</td>
<td>120</td>
<td>20.0%</td>
<td>61.7%</td>
</tr>
<tr>
<td>Sarah</td>
<td>3442</td>
<td>230</td>
<td>28.7%</td>
<td>76.8%</td>
</tr>
<tr>
<td>Naomi</td>
<td>1797</td>
<td>192</td>
<td>32.3%</td>
<td>61.9%</td>
</tr>
<tr>
<td>Peter</td>
<td>2112</td>
<td>139</td>
<td>25.9%</td>
<td>79.8%</td>
</tr>
<tr>
<td>Nina</td>
<td>2830</td>
<td>191</td>
<td>34.0%</td>
<td>77.2%</td>
</tr>
</tbody>
</table>
An artificial example

- Let there be 100 stems, all can take affix A, but only 90 take affix B: 10 are exceptions
- Assume the stem frequencies are Zipfian and the affix frequencies are also Zipfian
- Combine stems with affixes 1000 times according to the stem and suffix probabilities
- Count the affix overlap (% of stems that take both A and B over those that take either) and compare with the expected overlap if A and B were fully interchangable
  - empirical value should be lower than theoretical calculation
- Do this 100 times
• A good test that fails ...
Beyond Determiners

- Even very large samples of adult speech data show island-like pattern: few arguments are common, vast majority are rare
  - For “islands” to disappear requires billions of words
- Zipf-like distributions in morphology (Chan 2008, Chan & Lignos 2011)
  - When sample sizes and the number of stems are taken into account, child and adult morphology diversities from Italian, Spanish, and Catalan are largely the same
Conclusion

• The role of memory and usage as a replacement for grammar, even morphology, has been exaggerated.

• The child’s early grammar appears productive
  
  • She must be ready to generalize right away on data with relatively little diversity of usage (many tokens of few types)

• Other statistical tests for grammar are under development

• Matches vs. Mismatches in theoretical and experimental research