Usage Unevenness in Child Language Supports Grammar Productivity

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

BUCLD 2011
Saying and Knowing

★ What one says might not know reflect what knows about language

★ Bellugi, Bloom, Bowerman, Brown, Cazden, C. Chomsky, N. Chomsky, Fraser, McNeill, Schlesinger, Slobin

★ 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.
A simple observation

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

★ From the Harvard children

★ 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)

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

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

★ Brown corpus (Kucera & Francis 1967): 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)

★ allows one to approximate frequencies of words without even knowing what they are
The Grammar Hypothesis

★ Assume DP⇒DN is completely productive: combination is independent

★ D⇒a/the, N⇒cat, book, desk, water, dog ...

★ 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)
★ The $i$th word has probability of $P_r$.

\[
\frac{C/r}{\frac{C}{1} + \frac{C}{2} + \cdots + \frac{C}{N}}
\]

\[
\frac{1}{rH_N} \quad \text{where} \quad H_N = \sum_{i=1}^{N} \frac{1}{i}
\]

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

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

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

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

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

\[ \text{D=2, N=100, S=200} \]

math details: Yang (2011) ACL
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 (gpostttl.sourceforge.net): sufficiently adequate due to the unambiguity of “a” and “the”

★ extract D-N$_{sg}$ pairs
Empirical and Theoretical Results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sample Size (S)</th>
<th><em>a</em> or <em>the</em> Noun types (N)</th>
<th>Overlap (expected)</th>
<th>Overlap (empirical)</th>
<th>S/N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naomi (1;1-5;1)</td>
<td>884</td>
<td>349</td>
<td>21.8</td>
<td>19.8</td>
<td>2.53</td>
</tr>
<tr>
<td>Eve (1;6-2;3)</td>
<td>831</td>
<td>283</td>
<td>25.4</td>
<td>21.6</td>
<td>2.94</td>
</tr>
<tr>
<td>Sarah (2;3-5;1)</td>
<td>2453</td>
<td>640</td>
<td>28.8</td>
<td>29.2</td>
<td>3.83</td>
</tr>
<tr>
<td>Adam (2;3-4;10)</td>
<td>3729</td>
<td>780</td>
<td>33.7</td>
<td>32.3</td>
<td>4.78</td>
</tr>
<tr>
<td>Peter (1;4-2;10)</td>
<td>2873</td>
<td>480</td>
<td>42.2</td>
<td>40.4</td>
<td>5.99</td>
</tr>
<tr>
<td>Nina (1;11-3;11)</td>
<td>4542</td>
<td>660</td>
<td>45.1</td>
<td>46.7</td>
<td>6.88</td>
</tr>
<tr>
<td>First 100</td>
<td>600</td>
<td>243</td>
<td>22.4</td>
<td>21.8</td>
<td>2.47</td>
</tr>
<tr>
<td>First 300</td>
<td>1800</td>
<td>483</td>
<td>29.1</td>
<td>29.1</td>
<td>3.73</td>
</tr>
<tr>
<td>First 500</td>
<td>3000</td>
<td>640</td>
<td>33.9</td>
<td>34.2</td>
<td>4.68</td>
</tr>
<tr>
<td>Brown corpus</td>
<td>20650</td>
<td>4664</td>
<td>26.5</td>
<td>25.2</td>
<td>4.43</td>
</tr>
</tbody>
</table>

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

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

\[
S \frac{1}{r H_N} > 1
\]

\[
r = \frac{S}{H_N} \approx \frac{S}{\ln N}
\]

★ Overlap is a monotonically increasing function of

\[
\propto \frac{S}{N \ln N} \quad \text{or} \quad \frac{S}{N} \quad \text{as } \ln N \text{ grows slowly}
\]
### Analysis of Variation

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sample Size ($S$)</th>
<th>$a$ or the Noun types ($N$)</th>
<th>Overlap (expected)</th>
<th>Overlap (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naomi (1;1-5;1)</td>
<td>884</td>
<td>349</td>
<td>21.8</td>
<td>19.8</td>
</tr>
<tr>
<td>Eve (1;6-2;3)</td>
<td>831</td>
<td>283</td>
<td>25.4</td>
<td>21.6</td>
</tr>
<tr>
<td>Sarah (2;3-5;1)</td>
<td>2453</td>
<td>640</td>
<td>28.8</td>
<td>29.2</td>
</tr>
<tr>
<td>Adam (2;3-4;10)</td>
<td>3729</td>
<td>780</td>
<td>33.7</td>
<td>32.3</td>
</tr>
<tr>
<td>Peter (1;4-2;10)</td>
<td>2873</td>
<td>480</td>
<td>42.2</td>
<td>40.4</td>
</tr>
<tr>
<td>Nina (1;11-3;11)</td>
<td>4542</td>
<td>660</td>
<td>45.1</td>
<td>46.7</td>
</tr>
<tr>
<td>First 100</td>
<td>600</td>
<td>243</td>
<td>22.4</td>
<td>21.8</td>
</tr>
<tr>
<td>First 300</td>
<td>1800</td>
<td>483</td>
<td>29.1</td>
<td>29.1</td>
</tr>
<tr>
<td>First 500</td>
<td>3000</td>
<td>640</td>
<td>33.9</td>
<td>34.2</td>
</tr>
<tr>
<td>Brown corpus</td>
<td>20650</td>
<td>4664</td>
<td>26.5</td>
<td>25.2</td>
</tr>
</tbody>
</table>

$$r = 0.986, \ p<0.00001$$
Does memory+retrieval work?

“(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 evaluation: model the learner as a list of joint DN pairs with their associated frequency rather than independently combined units

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

**small 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)

★ Does not exhaust item/usage-based approaches
A Quick Look at Verbs

★ Zipf-like distributions in words, morphology and syntactic rules (Chan 2008, Chan & Lignos 2011)

★ Islands everywhere! (Kowalski & Yang, yesterday)

★ 1.1 million child-directed English sentences

★ Top 15 more frequent transitive verbs

★ Top 10 most frequent frames following Tomasello (1992)
  ★ e.g., “see him” and “see her”
# Zipf all the way

<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
</tr>
</thead>
<tbody>
<tr>
<td>put</td>
<td>401</td>
<td>164</td>
<td>124</td>
<td>15</td>
<td>12</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>tell</td>
<td>245</td>
<td>65</td>
<td>49</td>
<td>49</td>
<td>45</td>
<td>36</td>
<td>22</td>
<td>16</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>see</td>
<td>152</td>
<td>100</td>
<td>38</td>
<td>32</td>
<td>28</td>
<td>21</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>want</td>
<td>158</td>
<td>83</td>
<td>36</td>
<td>24</td>
<td>19</td>
<td>15</td>
<td>13</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>let</td>
<td>238</td>
<td>38</td>
<td>32</td>
<td>23</td>
<td>22</td>
<td>17</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>give</td>
<td>115</td>
<td>92</td>
<td>59</td>
<td>32</td>
<td>31</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>take</td>
<td>130</td>
<td>57</td>
<td>30</td>
<td>21</td>
<td>18</td>
<td>15</td>
<td>14</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>show</td>
<td>100</td>
<td>34</td>
<td>27</td>
<td>21</td>
<td>19</td>
<td>17</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>got</td>
<td>58</td>
<td>37</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>ask</td>
<td>45</td>
<td>41</td>
<td>27</td>
<td>24</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>make</td>
<td>67</td>
<td>20</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>eat</td>
<td>67</td>
<td>42</td>
<td>14</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>like</td>
<td>39</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>bring</td>
<td>43</td>
<td>30</td>
<td>17</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>hear</td>
<td>46</td>
<td>22</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>total</td>
<td>1904</td>
<td>838</td>
<td>501</td>
<td>301</td>
<td>252</td>
<td>189</td>
<td>137</td>
<td>109</td>
<td>88</td>
<td>75</td>
</tr>
</tbody>
</table>
Islands Everywhere

★ 100 verbs, 100 nouns: **10 million words** for 50% diversity

★ 1500 verbs, 1500 nouns: **46 years** for 50% diversity

---

And that is exactly the point: the advocates of item-based learning not only rejected the alternative hypothesis without adequate statistical tests but also accepted the favored hypothesis without adequate statistical validation. Intuition is no substitute for theoretical analysis or statistical validation.

### An Itemized Look at Verbs

The formal analysis in section 1 can be generalized to the study of child verb syntax and morphology. Unfortunately, the acquisition data in support of the Verb Island Hypothesis and the item-based nature of early morphology is not available in the public domain. But there is no escape from the Zipfian grasp: the combinatorics of verbs and their morphological and syntactic associates are similarly lopsided in their usage distribution as in the case of determiners.

Consider first the kind of verbal syntax distributions attributed to the Verb Island Hypothesis. We focus on constructions that involve a transitive verb and its nominal objects, including pronouns and noun phrases. Following the definition of "sentence frame" in Tomasello's original Verb Island study, each unique lexical item in the object position counts as a unique construction for the verb. Figure 1 shows the construction frequencies of the top 10 transitive verbs in sp million child-directed utterances.

![Graph showing rank and frequency of verb-object constructions based on sp million child-directed utterances.](image)

*Figure 1: Rank and frequency of verb-object constructions based on sp million child-directed utterances.*

Processing methods are as described in Box 1 except here we focus on adjacent verb-nominal pairs in part-of-speech tagged text. The verbs are the top 10 most frequent transitive verbs: *put, tell, see, want, let, give, take, show, got, ask, make eat, like, bring, and hear.* For each verb, we counted its top 10 most frequent constructions, which are defined as the verb followed by a unique lexical item in the object position. For example, "ask him" and "ask John" are different constructions. For each of the 10 ranks, we tallied the construction frequencies for all 10 verbs. The verbs are: *put, tell, see, want, let, give, take, show, got, ask, make eat, like, bring, and hear.*
Conclusion

★ Grammar + Zipf = Usage

★ One of the many (future) statistical tests for language

★ Child language: Is there a storage stage? (Possible, but let’s catch it early!)

★ Productivity is not inconsistent with storage effects

★ Theory of grammar: The role of storage in syntactic coverage is minimal (Bikel 2004, *Computational Linguistics*)

★ Matches vs. mismatches in theoretical and experimental research

★ The most important lesson from Zipf...
WE ARE THE 99%