Usage Uneveness 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

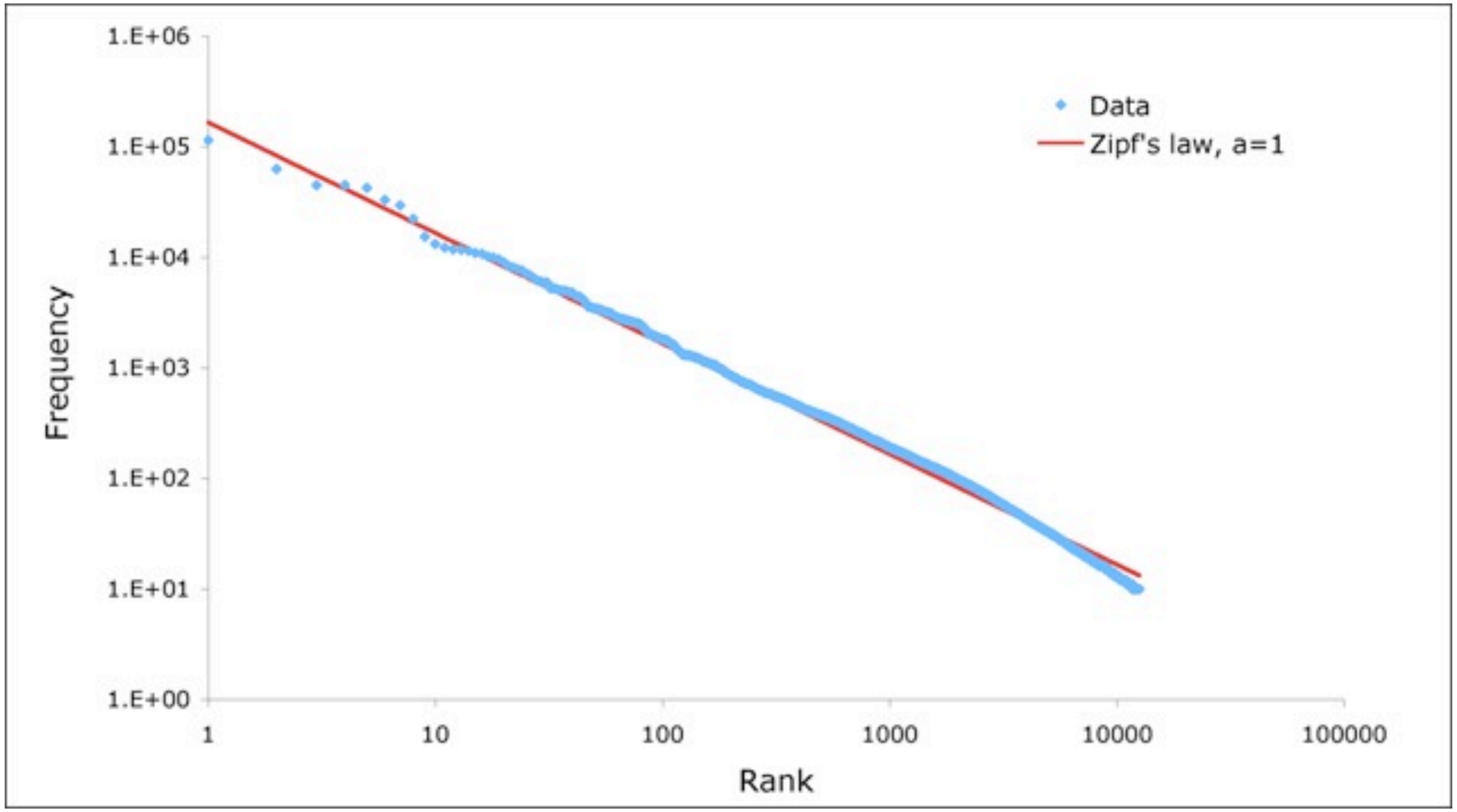
- ★Valian (1986): the knowledge of the category **determiner** fully productive by 2;0, virtually no errors
    - ★low error rate could be memory and retrieval
- ★Pine & Lieven (1997): **overlap** is much lower than, say, even 50% (following Tomasello's verb island hypothesis)

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

- ★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

$$\text{rank} = \frac{\text{C}}{\text{frequency}} \qquad \log(\text{rank}) = \log C - \log(\text{frequency})$$



- ★ 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**)

Zipfian Probabilities

★ The *r*th word has **probability** of $\mathbf{P_r}$

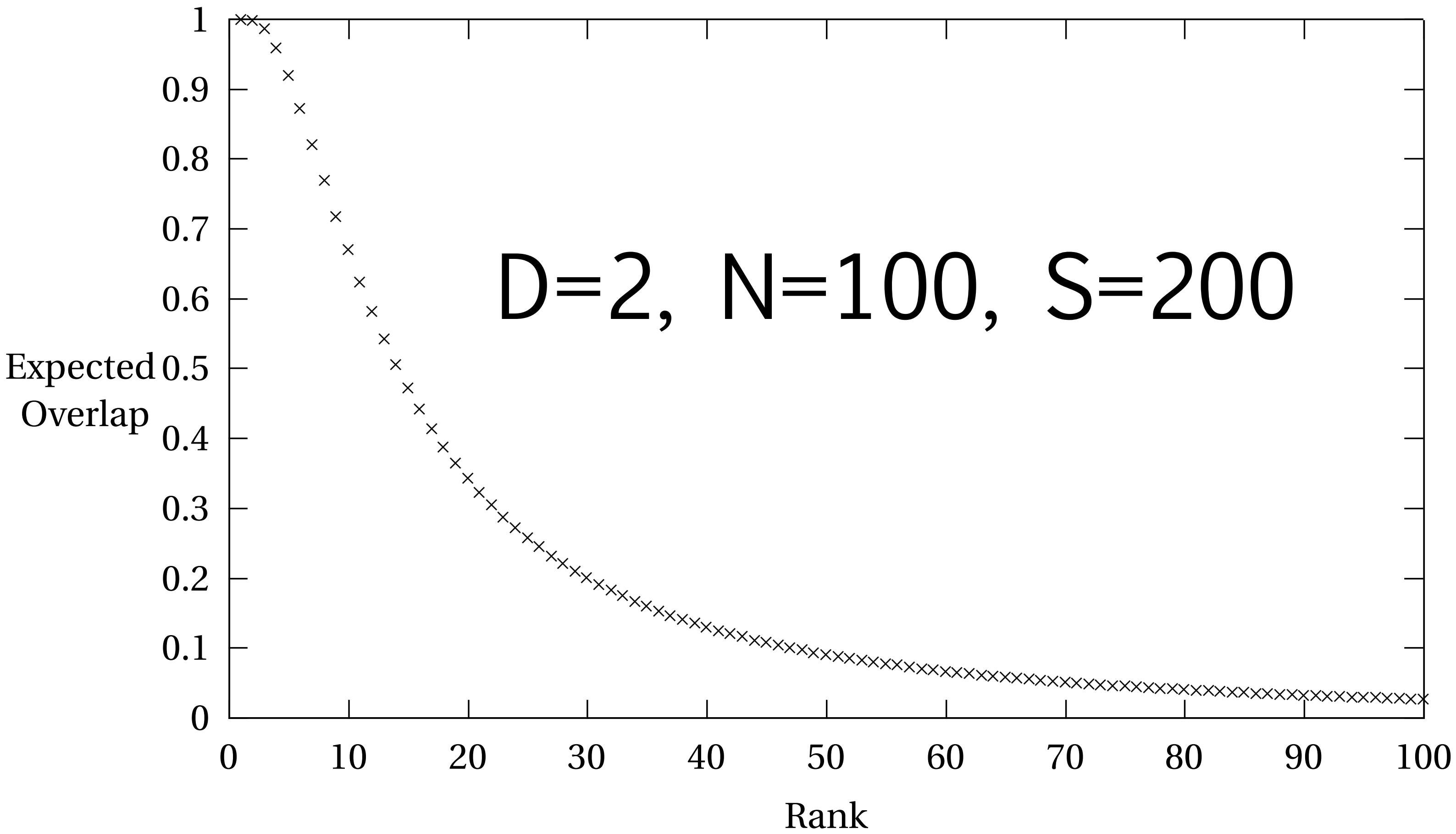
$$\frac{C/r}{\frac{C}{1}+\frac{C}{2}+...+\frac{C}{N}}$$

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

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

Expected overlap

$$\begin{aligned} O(n_r) &= 1 - \Pr\{n_r \text{ is not sampled during } S \text{ trials}\} \\ &- \sum_{i=1}^{D} \Pr\{n_r \text{ is sampled but with the } i\text{th determiner exclusively}\} \\ &= 1 - (1-p_r)^S \\ &- \sum_{i=1}^{D} \Big[(d_i p_r + 1 - p_r)^S - (1-p_r)^S \Big] \end{aligned}$$



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 (gposttl.sourceforge.net): sufficiently adequate due to the unambiguity of “**a**” and “**the**”
- ★ extract D-$N_{\mathbf{sg}}$ pairs

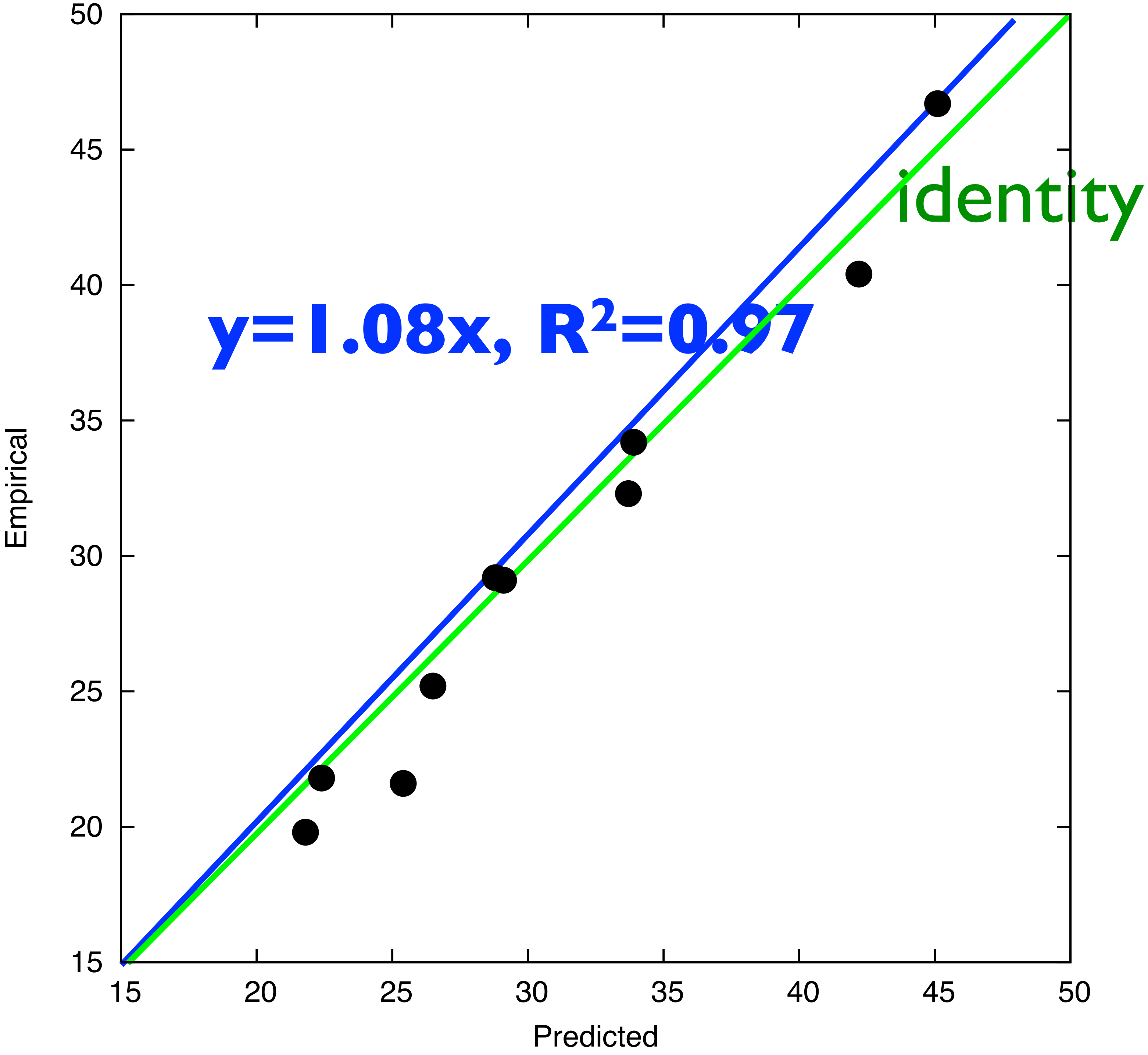
Empirical and Theoretical Results

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

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



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

$$\propto \frac{S}{N\ln N} \text{ or } \frac{S}{N} \text{ as } \ln N \text{ grows slowly}$$

Analysis of Variation

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

$r = 0.986$, $p < 0.0001$

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

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

★paired t- and Wilcoxon tests show significant differences ($p < 0.005$)

★Does not exhaust item/usage-based approaches

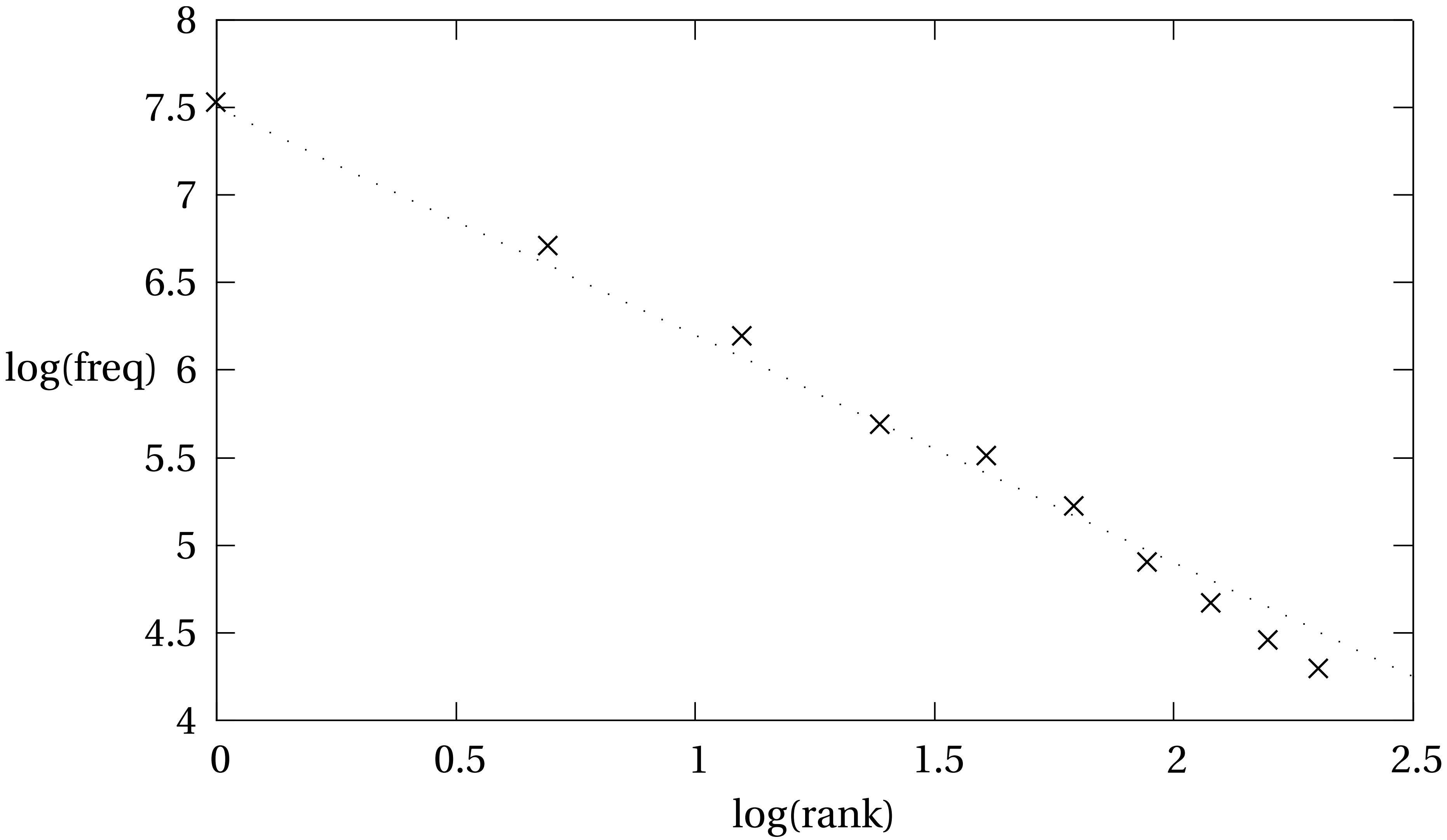
A Quick Look at Verbs

- ⋆ 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

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

Islands Everywhere



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

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

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%