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)

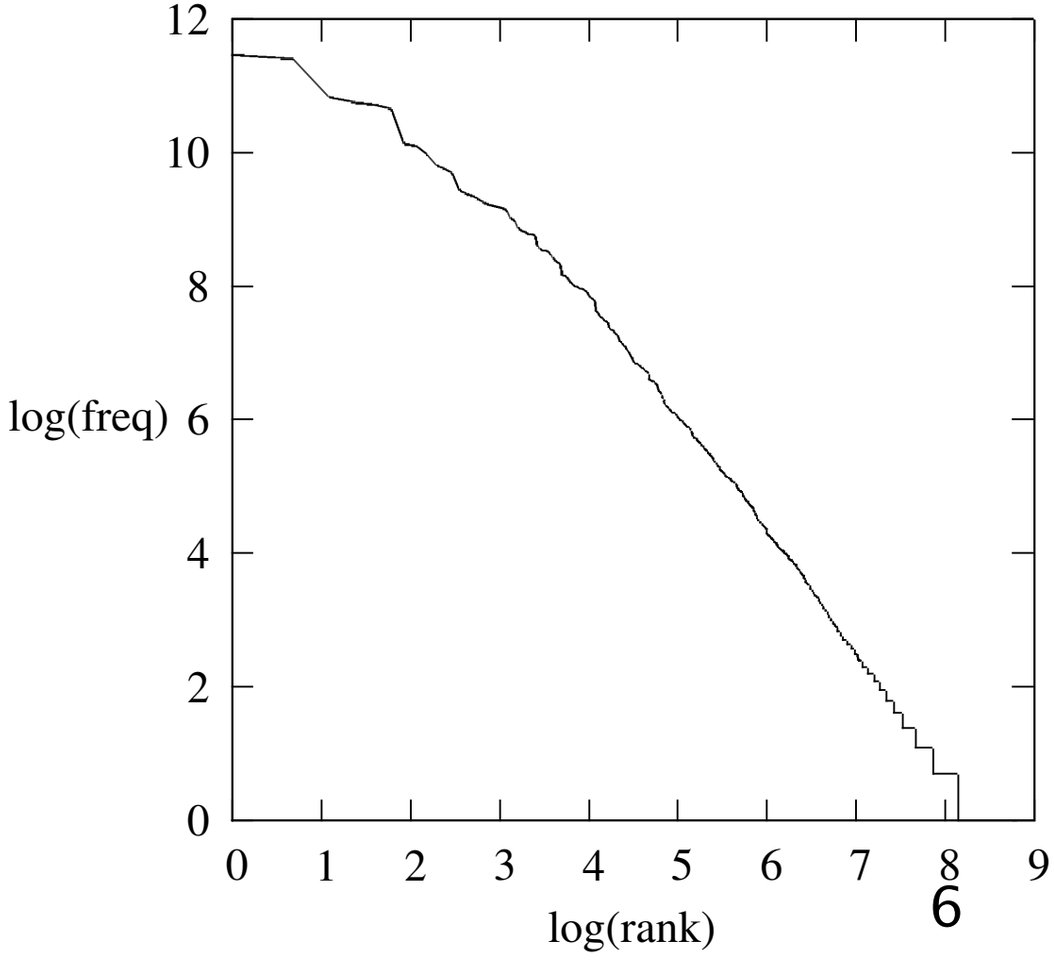
$$\text{overlap} = \frac{\#\text{ of nouns with BOTH } \textit{the} \text{ AND } \textit{a}}{\#\text{ of nouns with EITHER } \textit{the} \text{ OR } \textit{a}}$$

  - 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

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

- 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



courtesy: Constantine Lignos

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 $\mathbf{P_r}$

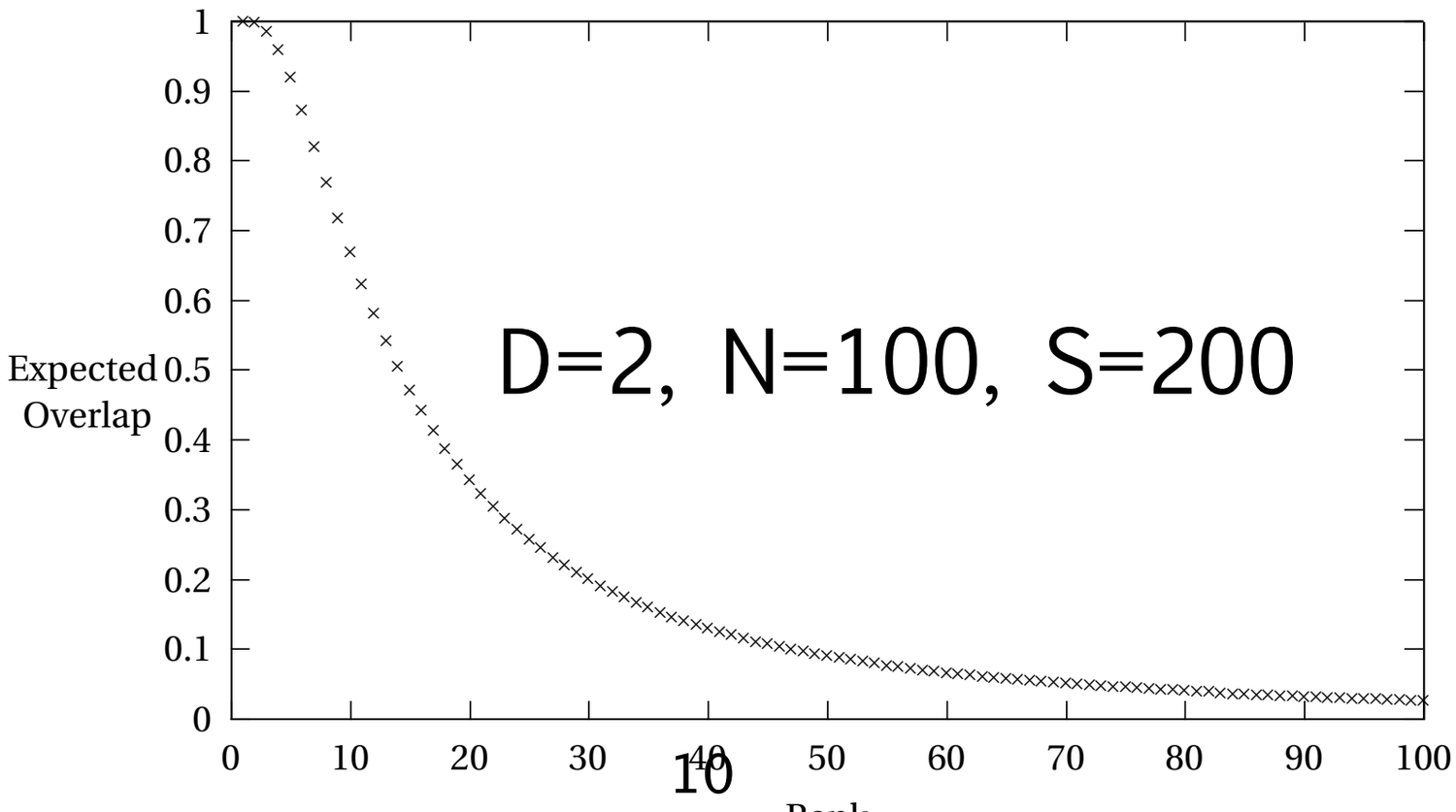
$$\frac{C/r}{\frac{C}{1}+\frac{C}{2}+\cdots+\frac{C}{N}}$$

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

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

Expected overlap

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



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**"
- Standard procedure in child data processing:
  - remove annotation markers
  - repetitions count only once ("a doggie! a doggie! a doggie!")
  - extract D-N_{sg} pairs

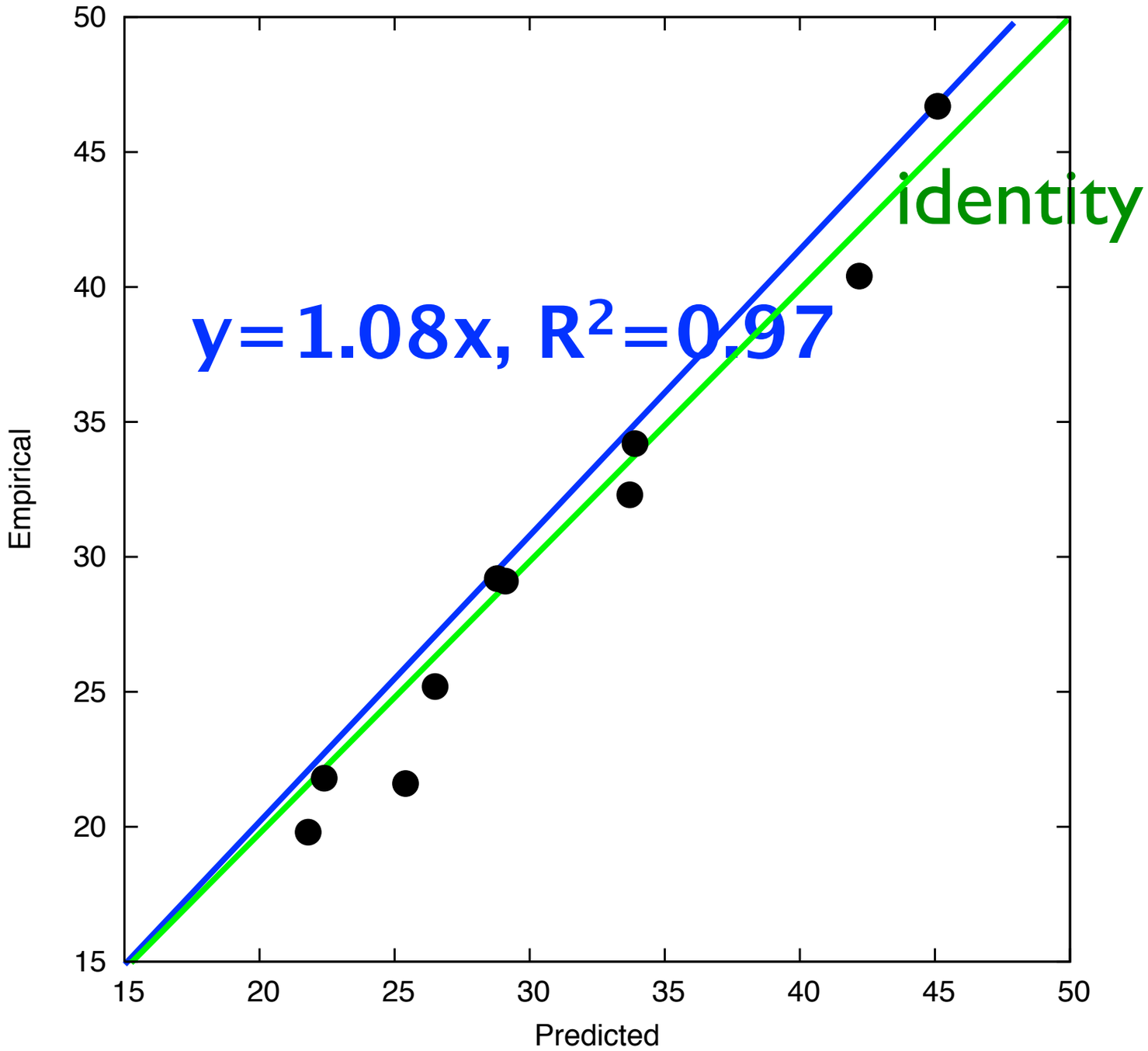
Empirical and Theoretical Results

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



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

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

$$r = 0.986,\ p<0.00001$$

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

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 ($p < 0.005$)

Measuring Productivity

- The calculation should distinguish productive and unproductive processes as measured by interchangeability
- An example in morphology
  - V_{INFL} → V + suffix
  - suffix → 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

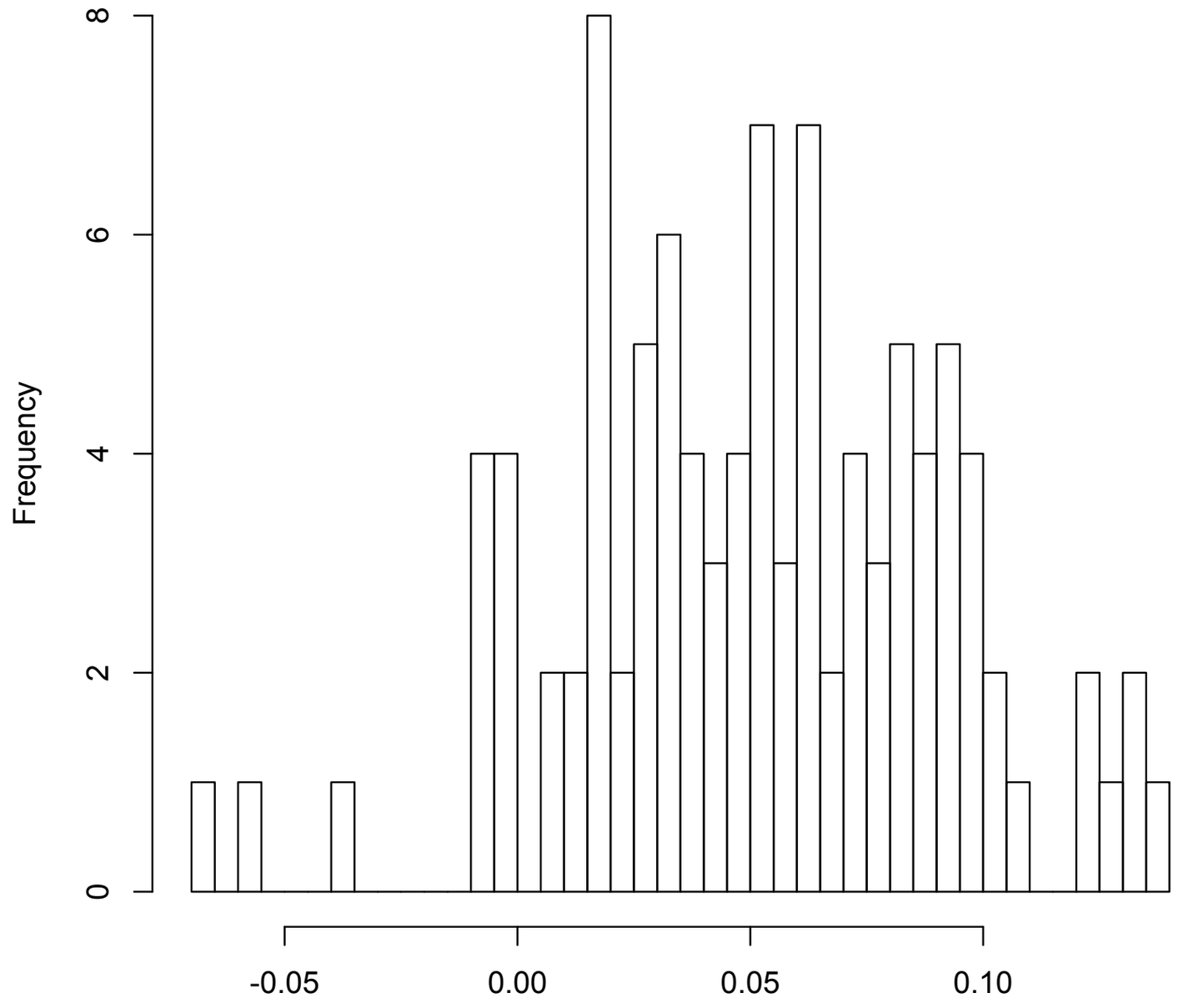
-ed vs. -ing

File	S	# V	Emp.	Theo.
Brown	62807	3044	**45.5%**	**75.6%**
Adam	6774	263	**31.3%**	**90.5%**
Eve	1028	120	**20.0%**	**61.7%**
Sarah	3442	230	**28.7%**	**76.8%**
Naomi	1797	192	**32.3%**	**61.9%**
Peter	2112	139	**25.9%**	**79.8%**
Nina	2830	191	**34.0%**	**77.2%**

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 calcuation
- Do this 100 times

(Theoretical - Empirical)



- A good test that fails ...

Beyond Determiners

- Even very large samples of adult speech data shows 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
- Sparse data problem: diminishing returns of bilexicalized rules (= “frames”, Tomasello 1992) from statistical parsing literature (Gildea 2001, Klein & Manning 2003, Bikel 2004)

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