On Usage and Grammar

Charles Yang
Department of Linguistics & Computer Science
University of Pennsylvania

Input & Syntactic Acquisition, Irvine, CA
9-11-2009

Seeing vs. Believing

- Mismatch between the input and the output
- Different kinds of input do different things (Yang 2002, 2004, 2005; see also Fodor & Sakas 2002, Pearl 2007)
- Data, model, and inference, for both the child and the scientist
    - what's the grammar like based on the sample data?
    - what to do with the grammar given the sample data?

Usage/Item/Constructivist Language

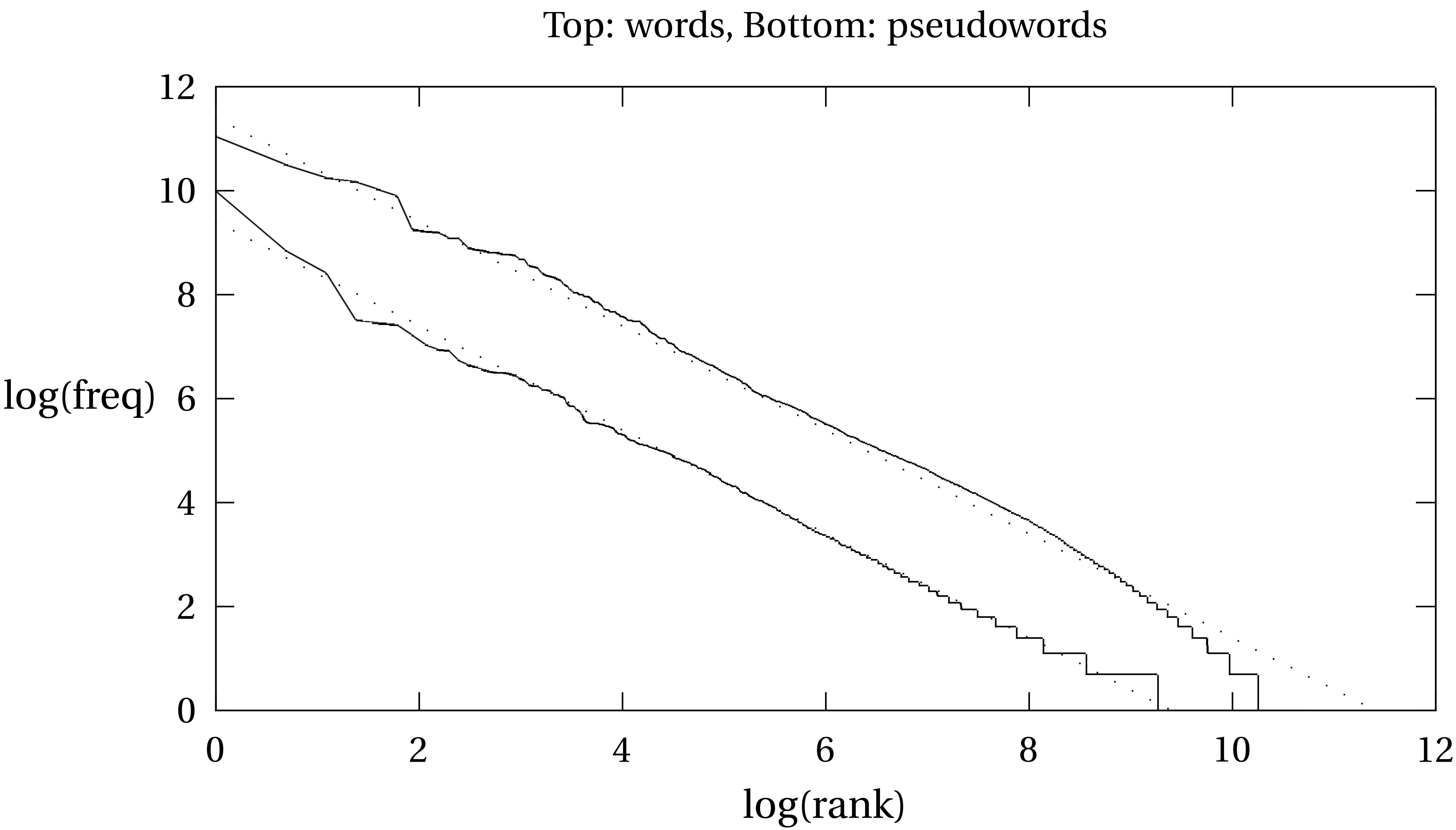
- “the 2-year-old child's syntactic competence is comprised totally of verb-specific constructions with open nominal slots. (Tomasello 2000, p214, *Cognition,* 2000, *TICS, etc.*)
    - **Verb Island Hypothesis** (Tomasello 1992): “Of the 162 verbs and predicate terms used, **almost half** were used in one and only one construction type, and **over two-thirds** were used in either one or two construction types.”
    - **Determiners** (Pine & Lieven 1997): far below chance overlap in “a-N” and “the-N” combinations, suggesting that determiner is not mastered early on (contra Valian 1986)
    - **morphology** (Pizutto & Caselli 1994): **47%** of all verbs were used in 1 person-number agreement (6 forms are possible), **40%** were used in 2 or 3 forms, and only **13%** were used in 4 or more.
- Statistical validations?

George Kingsley Zipf

- Few words are used frequently, and they are very frequent
- Most words are used very rarely, exactly only once
- More precisely, the rank and the frequency of words multiple to a constant
  - this can be visualized by plotting log(rank) against log (frequency): you'll get a straight line

$$f = \frac{C}{r} \text{ where } C \text{ is some constant}$$

Plotting the Brown corpus

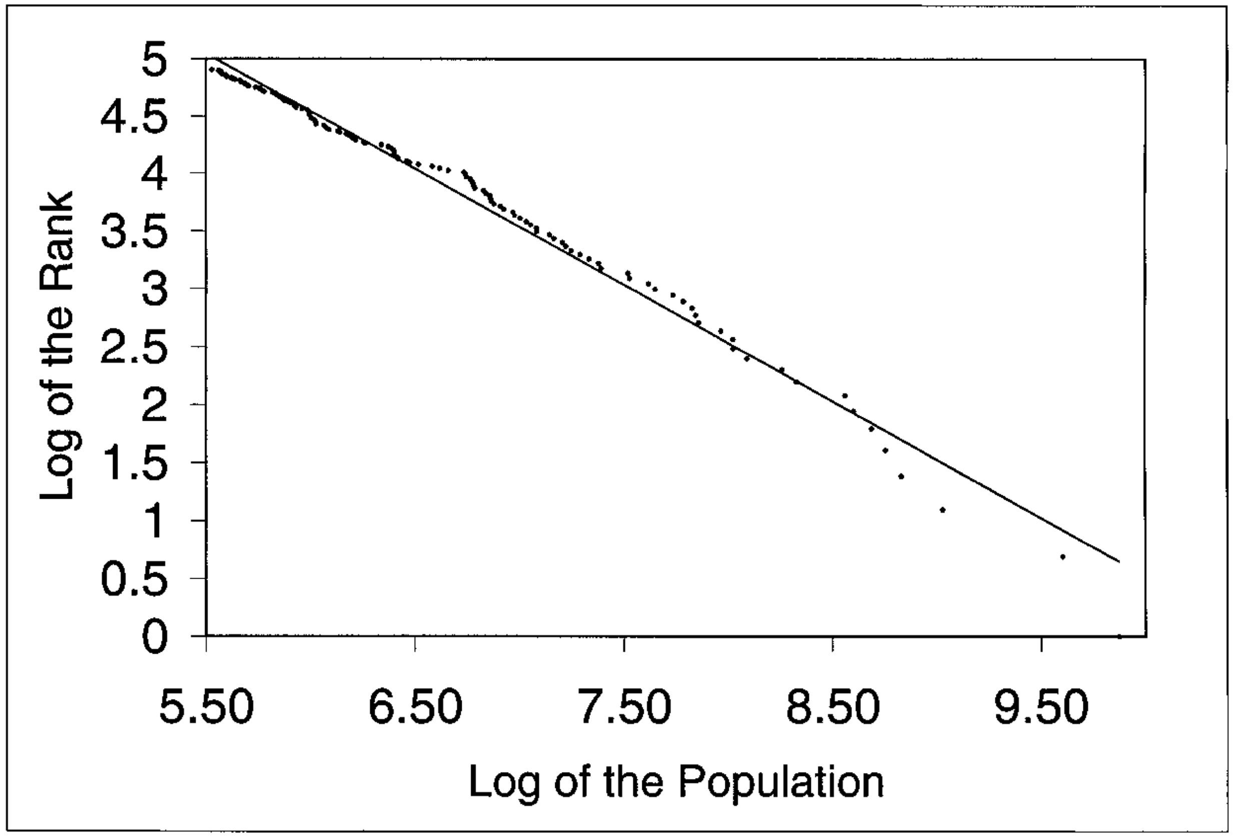


a perfect Zipf fit will have the slope of -1

But why?

- Zipf (1949): Principle of Least Effort (more frequent words tend to be shorter)
  - cf. the debate between Simon & Mandelbrot in the 1960s
- Chomsky (1958): define words as strings between, say, “e”
  - the_manager_is_late
  - th
  - _manag
  - r_is_lat

Zipfian presence



- Zipf like distributions can be observed all over the place
- In language, the slope of log-log fit is very close to 1 (Baroni 2008)

Usage/Item/Constructivist Language

- "the 2-year-old child's syntactic competence is comprised totally of verb-specific constructions with open nominal slots. (Tomasello 2000, p214, inter alia)
  - **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."
  - **Determiners** (Pine & Lieven 1997): far below chance overlap in "a-N" and "the-N" combinations, suggesting that determiner is not mastered early on (contra Valian 1986)
  - **morphology** (Pizutto & Caselli 1994): 47% of all verbs were used in 1 person-number agreement (6 forms are possible), 40% were used in 2 or 3 forms, and only **13%** were used in 4 or more.
- **But what's the Null Hypothesis?**

Diversity of Usage

- Valian (1986): the knowledge of the category **determiner** fully productive by 2;0, virtually no errors
- Pine & Lieven (1997), Pine & Martindale (1996): No, because **overlap** is much lower than, say, even 50%

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

  - The same logic behind Tomasello's Verb Island Hypothesis
  - 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%

The Productivity Hypothesis

- Assume DP DN is completely productive: combination is independent
    - D **a/the**, N **cat**, **book**, **desk**, …
    - substitute DP for VP, PP, inflections …
- 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

Determiner-Noun Usage

- “the bathroom” ≫ “a bathroom”
- “a bath” ≫ “the bath”
- Brown corpus: 75% of singular nouns occur with only **the** or **a**
  - **25%** of the remainders are balanced
  - 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**

Zipfian Probabilities

- Assume that there are N words and their frequencies are Zipfian
  - In the child production data, singular nouns have the slope of -1.08 (very close to perfect Zipfian fit)
- 1st word has frequency of C
- 2st word has frequency of C/2
- …
- rth word has frequency of C/r
- …

Zipfian Probabilities

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

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

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

- In a sample size of **S**, it has an expected occurrence of

$$SP_r = \frac{S}{rH_N}$$

- S and N can be directly obtained from CHILDES data

Expected overlap of the rth Noun

The expected overlap for N_r is 1 - (expected probability of N_r appearing with exactly one determiner for all SP_r trials), or

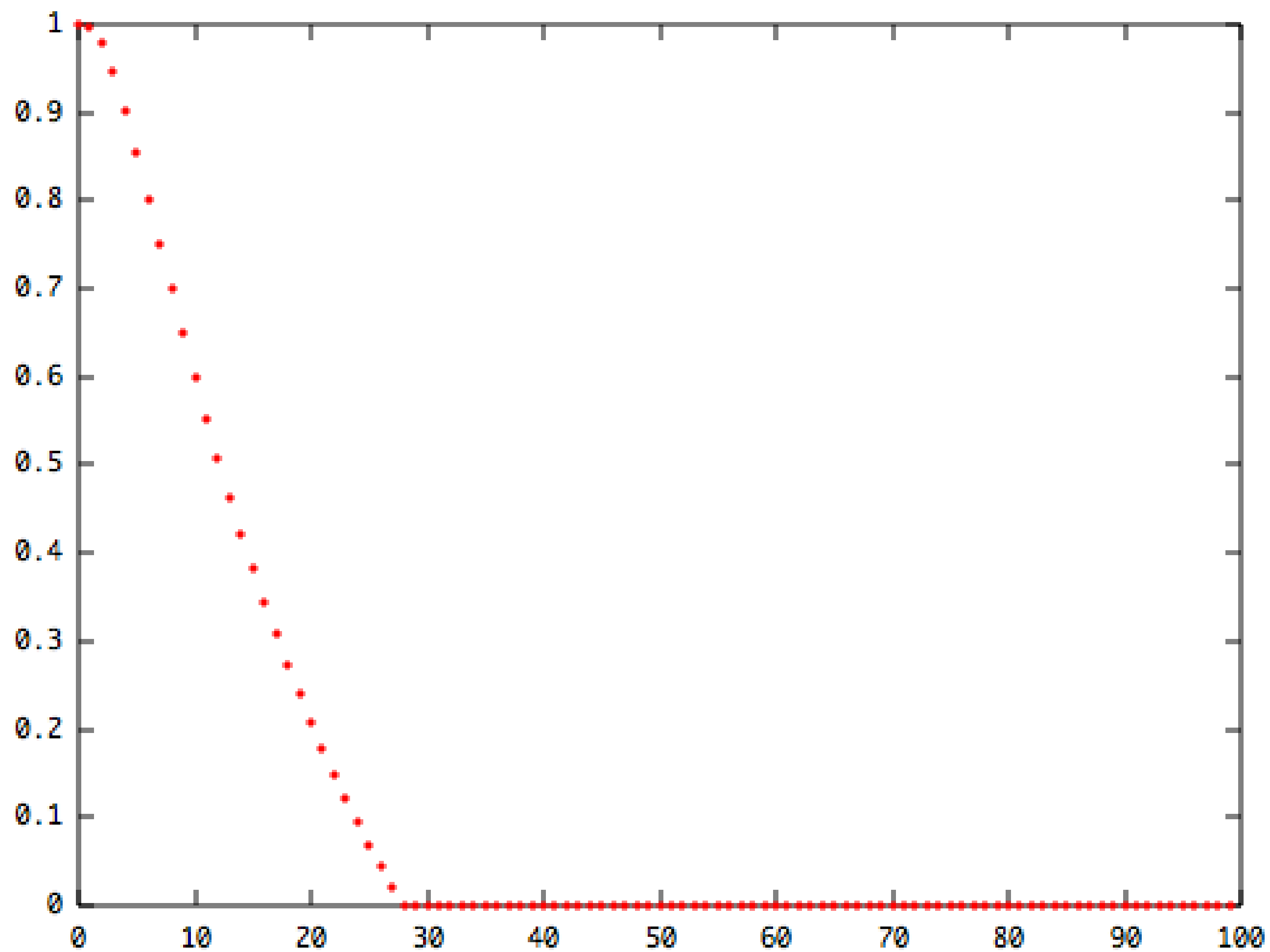
$$1 - \sum_{d=1}^{D} p_d^{SP_r} \tag{1}$$

If the determiners are also Zipfian, we have

$$p_d = \frac{1}{dH_D} \text{ where } H_D = \sum_{i=1}^{D} \frac{1}{d}$$

We add up (1) for all N nouns and divide that by N: that is expected overlap

N=50, S=100



- The expected probability of a noun having overlap

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**), which has an accuracy of about 97%
- Standard procedure in data processing:
  - remove annotation markers
  - repetitions count only once (“a doggie! a doggie! a doggie!”)
  - extract D-$N_{singular}$ pairs

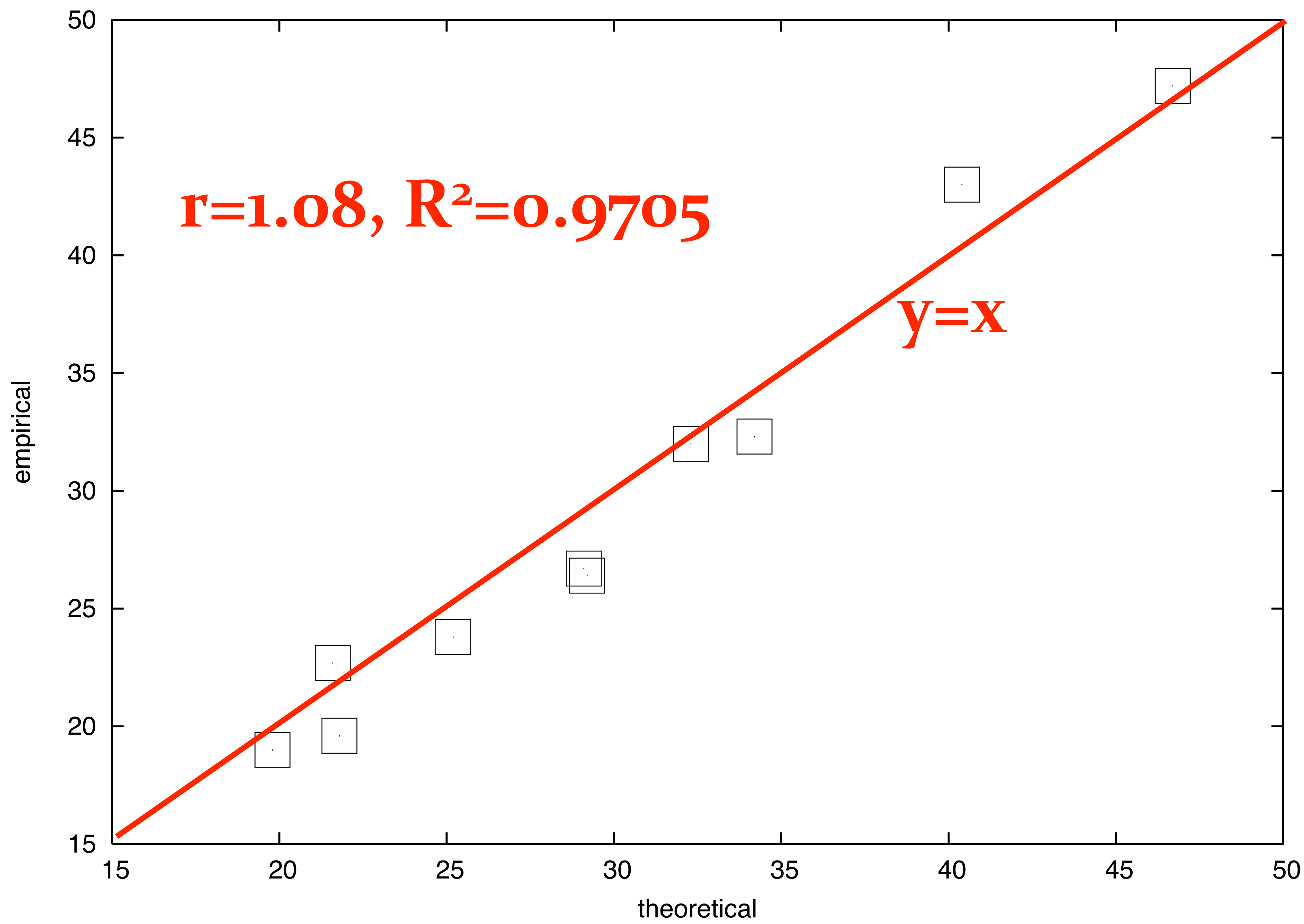
Empirical/Theoretical Results

Child	Sample Size (S)	*a* & *the* Noun types	*a* or *the* Noun types (N)	Overlap (expected)	Overlap (empirical)	$\frac{S}{N}$
Naomi (1;1-5;1)	884	60	349	19.0	19.8	2.53
Eve (1;6-2;3)	831	61	283	22.7	21.6	2.94
Sarah (2;3-5;1)	2453	187	640	26.4	29.2	3.83
Adam (2;3-4;10)	3729	252	780	32.0	32.3	4.78
Peter (1;4-2;10)	2873	194	480	43.0	40.4	5.99
Nina (1;11-3;11)	4542	308	660	47.2	46.7	6.88
First 100	600	53	243	19.6	21.8	2.47
First 300	1800	141	483	26.7	29.1	3.73
First 500	3000	219	640	32.3	34.2	4.68
Brown corpus	20650	1175	4664	23.8	25.2	4.43

also considered the first 100, 300, 500 tokens of the six children

paired t- and Wilcoxon tests reveal no difference

Empirical vs. Expected



Why Variation

- Some children have higher overlap than others (and Brown)
- 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

Child	Sample Size (S)	*a* & *the* Noun types	*a* or *the* Noun types (N)	Overlap (expected)	Overlap (empirical)	$\frac{S}{N}$
Naomi (1;1-5;1)	884	60	349	19.0	19.8	2.53
Eve (1;6-2;3)	831	61	283	22.7	21.6	2.94
Sarah (2;3-5;1)	2453	187	640	26.4	29.2	3.83
Adam (2;3-4;10)	3729	252	780	32.0	32.3	4.78
Peter (1;4-2;10)	2873	194	480	43.0	40.4	5.99
Nina (1;11-3;11)	4542	308	660	47.2	46.7	6.88
First 100	600	53	243	19.6	21.8	2.47
First 300	1800	141	483	26.7	29.1	3.73
First 500	3000	219	640	32.3	34.2	4.68
Brown corpus	20650	1175	4664	23.8	25.2	4.43

$r = 0.986, p < 0.00001$

Interim Conclusion

- Children's determiner usage is consistent with the hypothesis of fully productivity.
- We need a theory for how the child gets there (Yang 2002, 2005)
- 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

An attempt at item-based learning

- central tenet of frequency and memorization
  - model the learner as a list of **joint D-N** pairs with their associated frequency
  - sample from the list with their joint frequencies
- BIG learner: list consists of 1.1 million child-directed utterances
- small learner: list consists of the child-directed utterance for each particular child
- calculate the overlap for the sampled D-N pairs, averaging over 1000 trials

item-based learners

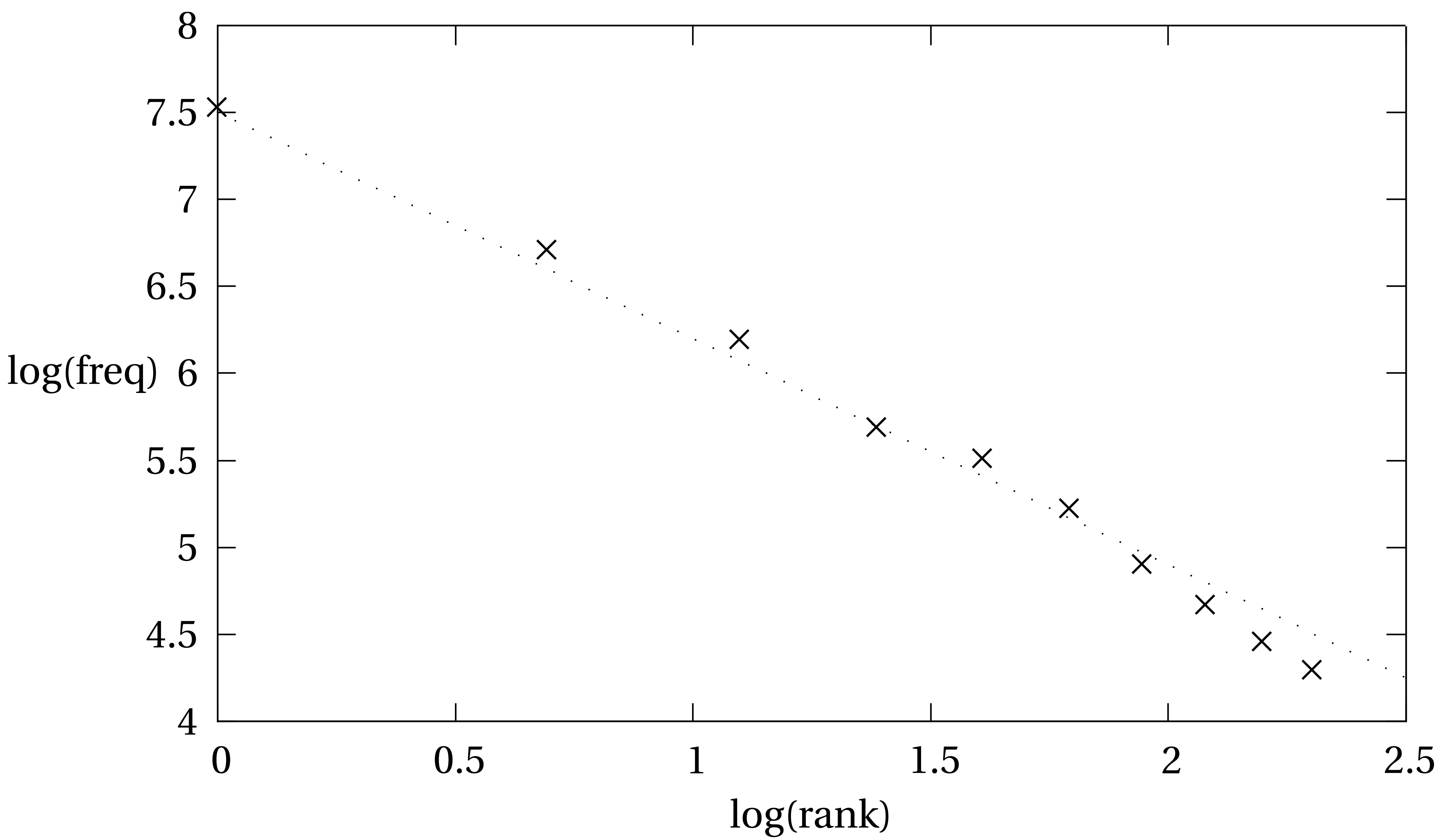
Child	Sample Size (S)	Overlap (BIG learner)	Overlap (small learner)	Overlap (empirical)
Eve	831	16.0	17.8	21.6
Naomi	884	16.6	18.9	19.8
Sarah	2453	24.5	27.0	29.2
Peter	2873	25.6	28.8	40.4
Adam	3729	27.5	28.5	32.3
Nina	4542	28.6	41.1	46.7
First 100	600	13.7	17.2	21.8
First 300	1800	22.1	25.6	29.1
First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

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

A brief look at verbs

- Tomasello (2000, p214, inter alia)
  - **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."
    - data not available in public
  - **morphology** (Pizutto & Caselli 1994): 47% of all verbs were used in 1 person-number agreement (6 forms are possible), 40% were used in 2 or 3 forms, and only **13%** were used in 4 or more.
    - data not available in public
- Data from CHILDES data

Islands all over the map …



- 1.1 million adult sentences, top 15 transitive verbs (*put, tell, see, want, let, give, take, show, got, ask, make, eat, like, bring, hear)*
- extracted top 10 "sentence frames" (Tomasello 1992) with nominal objects

Romance morphology

- (1, 2, 3 person) x (singular, plural) = 6 possible forms
- Data: entire Italian, Spanish, and Catalan child data and child-directed data
- Part-of-speech tagging preprocessing (**freeling**)
  - thanks to Erwin Chan for his help
- Only looking at finite forms (infinitives do not mark agreement)

Results

Subject	1 form	2 forms	3 forms	4 forms	5 forms	6 forms	token/type
Italian children	81.8%	7.7%	4.0%	2.5%	1.7%	0.3%	1.533
Italian adults	63.9%	11.0%	7.3%	5.5%	3.6%	2.3%	2.544
Spanish children	80.1%	5.8%	3.9%	3.2%	3.0%	1.9%	2.233
Spanish adults	76.6%	5.8%	4.6%	3.6%	3.3%	3.2%	2.607
Catalan children	69.2%	8.1%	7.6%	4.6%	3.8%	2.0%	2.098
Catalan adults	72.5%	7.0%	3.9%	4.6%	4.9%	3.3%	2.342

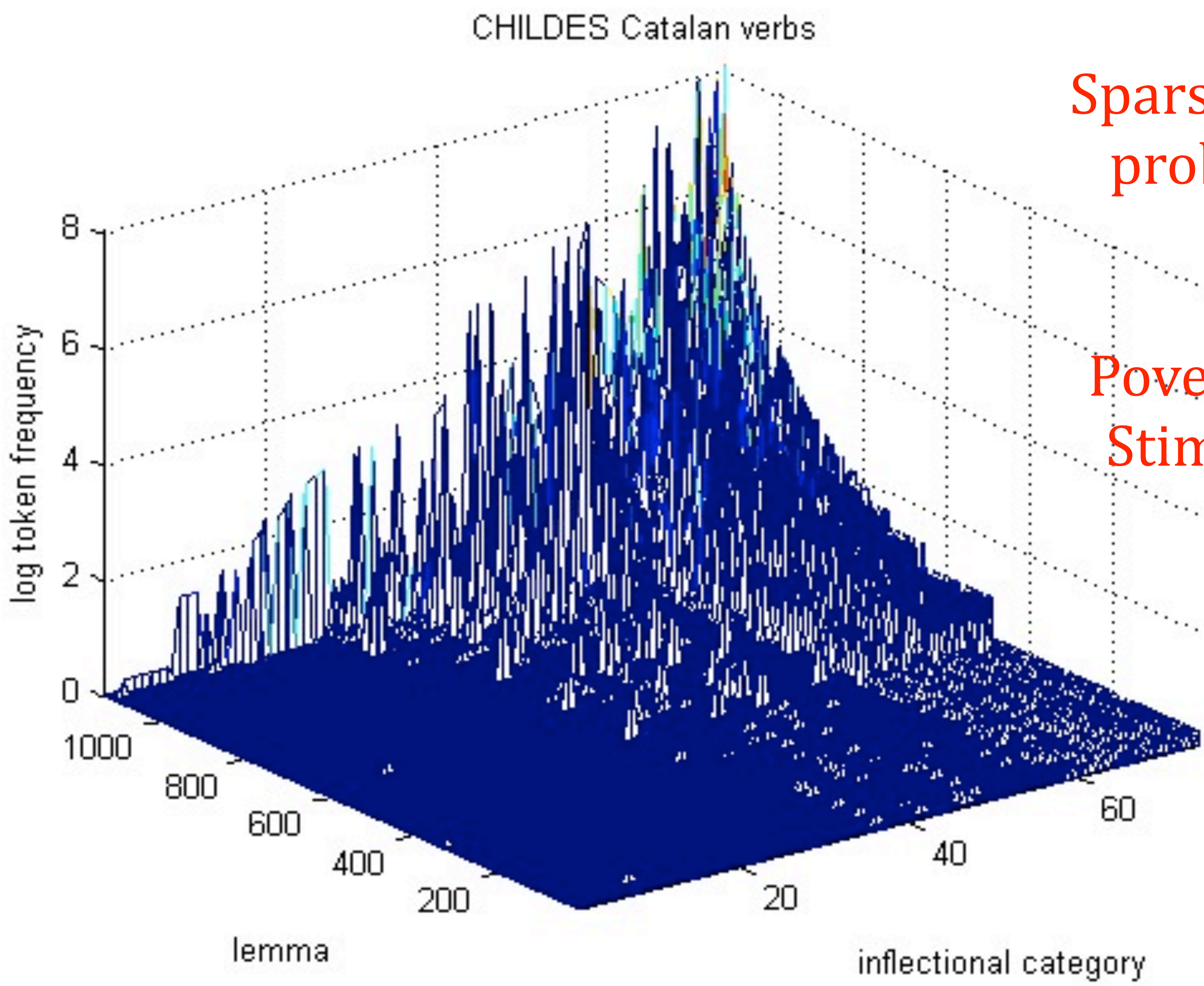
No major difference between Spanish & Catalan kids and adults

The main predictor is also S/N

A view from afar

- Sparse data problem (Jelinek 1993)
- The biggest challenge in computational linguistics
- Statistical models of language, even simple ones such as bigrams/trigrams, require parameter values: many will have zero occurrence in the sample, however large
- Sparse data problem = Poverty of Stimulus
- This is true at the level of morphology (Chan 2008)

Zipf Morph



Sparse data problem

=

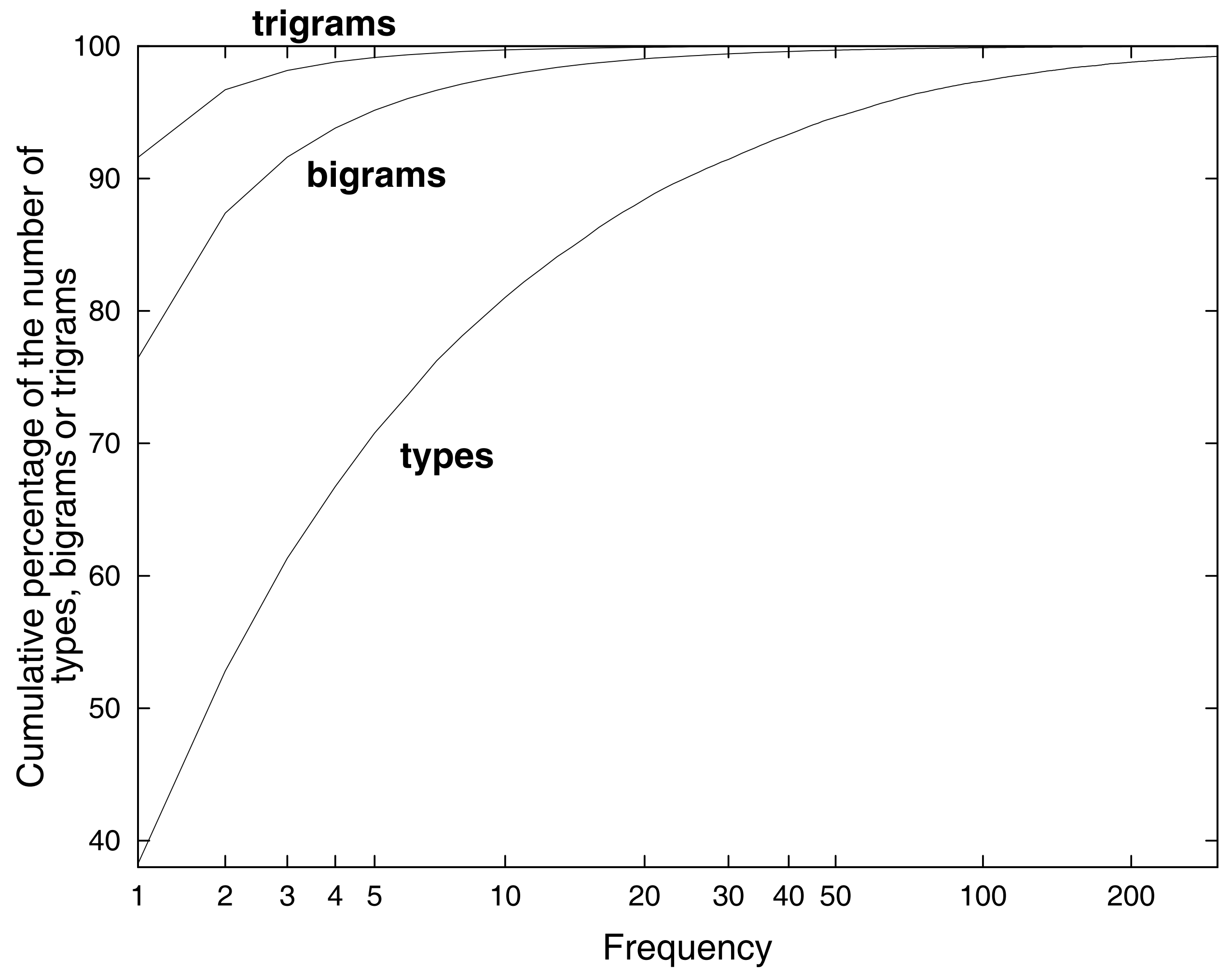
Poverty of Stimulus

No full paradigms in data

	Millions of words	# possible verb forms	Max # forms for any lemma
Basque	0.6	22	16
Catalan	1.7	45	33
Czech	2.0	72	41
Greek	2.6	83	45
Hungarian	1.0	76	48
Hebrew	2.5	33	23
Slovene	2.5	32	24
Spanish	2.7	51	34

From Chan (2008)

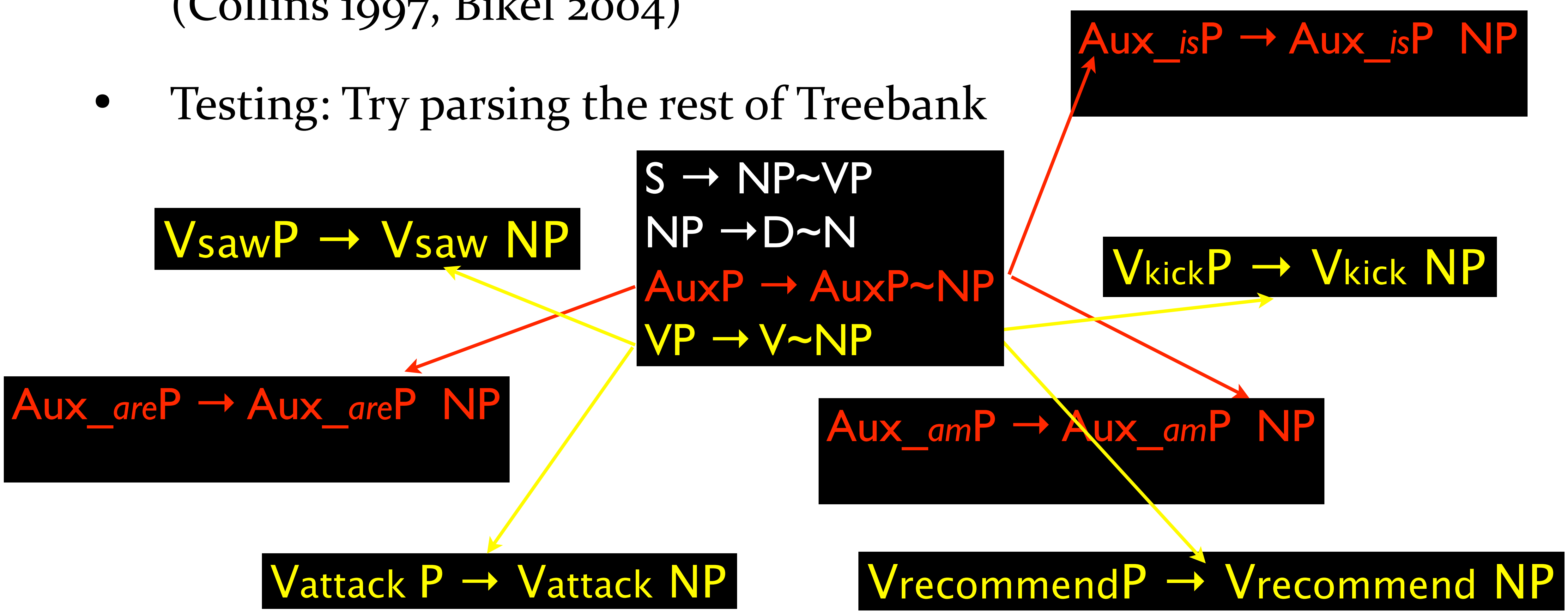
More Zipfian than Zipf



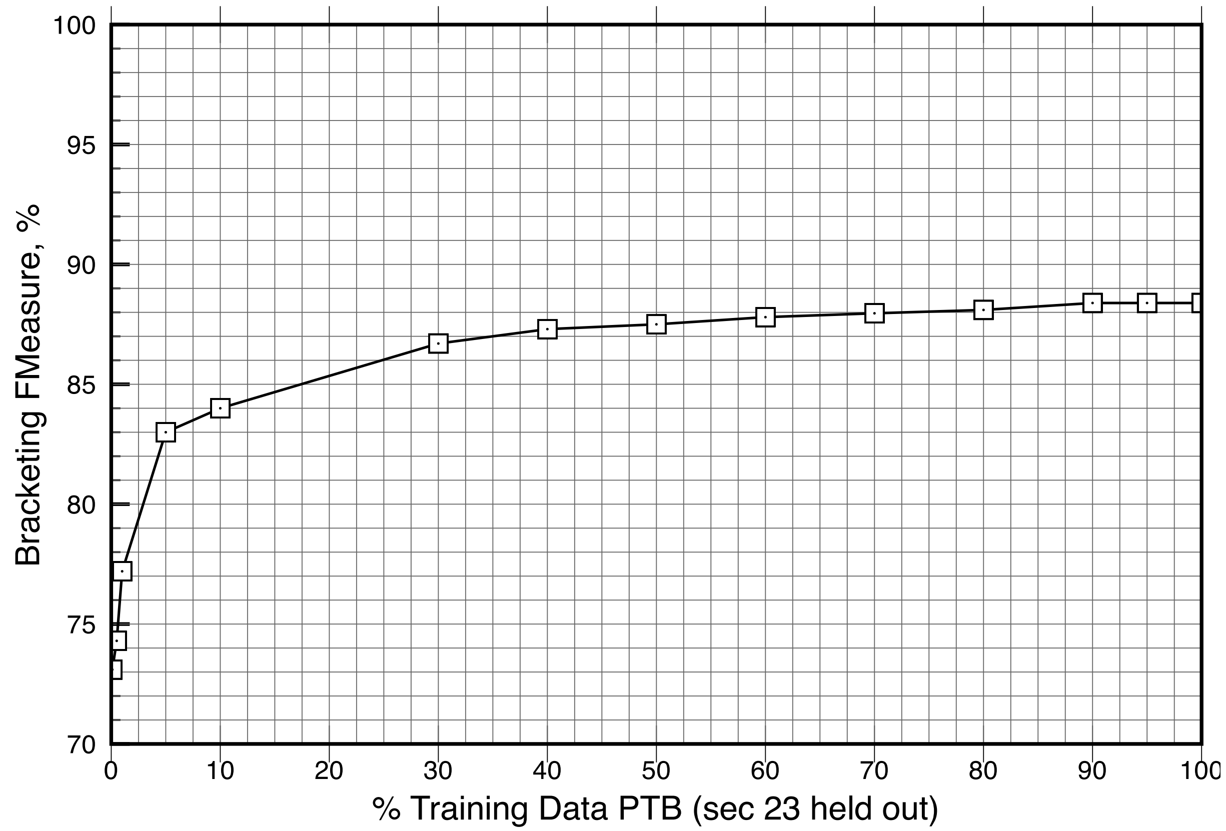
- Most things we hear are repetitions: this becomes more significant when we look at linguistic combinations
- Words, bigram, and trigram frequencies
    - 40% of words occur only once, ~80% of word-pairs occur only once, >90% of word triples occur only once

The view from (not too) afar

- Statistical language learning and parsing
- Input: most of Penn Treebank, a collection of trees manually annotated,
- Training: essentially tally the frequencies of usage for a predefined set of rules in the form of a **lexicalized** Context-Free Grammar (Collins 1997, Bikel 2004)
- Testing: Try parsing the rest of Treebank

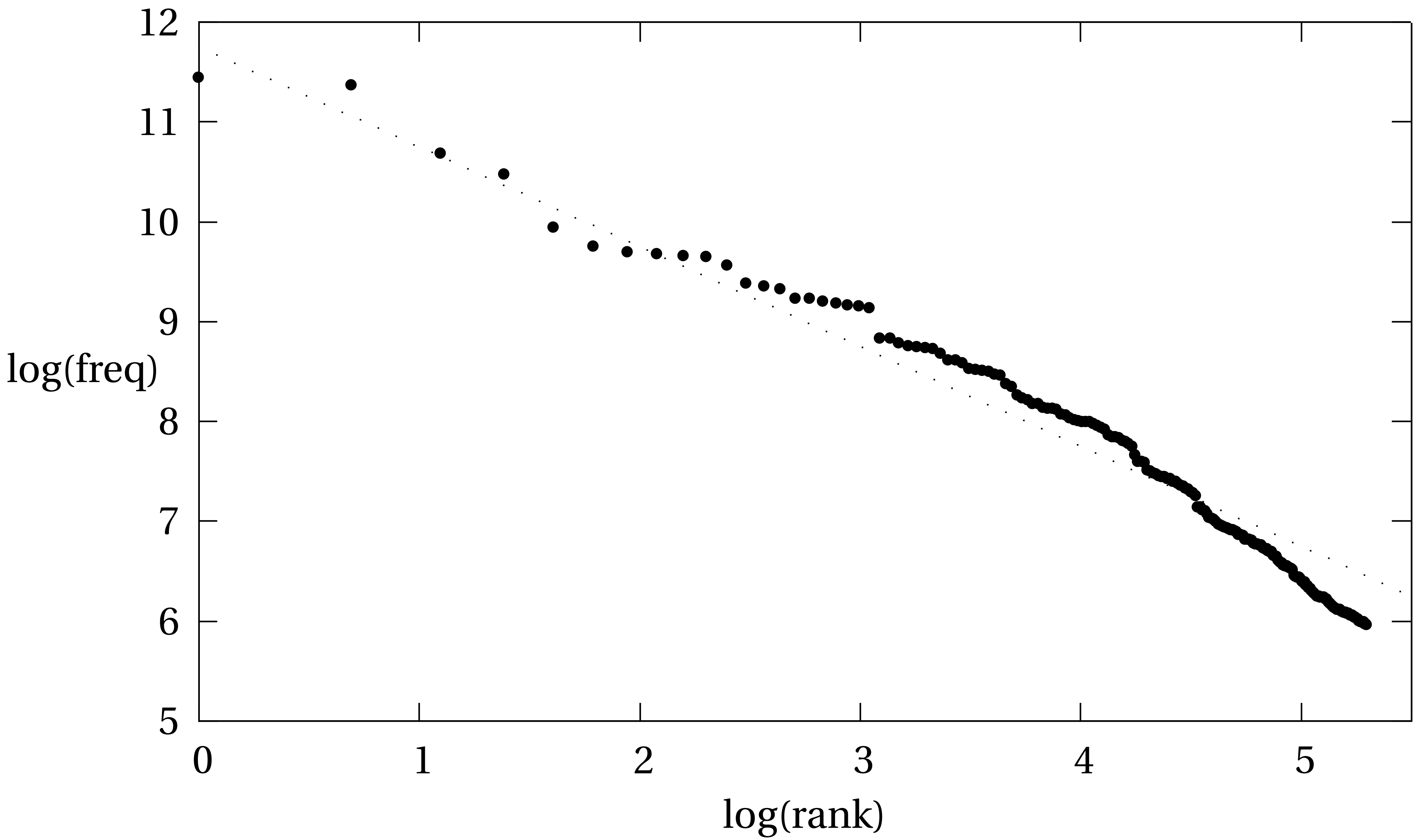


The law of diminishing returns



- The measurement of success is the % of phrase structure brackets inserted correctly
- The "learner" can quickly learn the general rules of the grammar with a small fraction of the data
- Lexicalized learning pays very little dividend (Bikel 2004)

Zipf in the Penn Treebank



- No. 1 rule: PP→ P NP
- No. 2 rule: S → NP VP

Grammar despite Usage

→

Grammar must overcome Usage