Who’s Afraid of George Kingsley Zipf? Charles Yang 2010

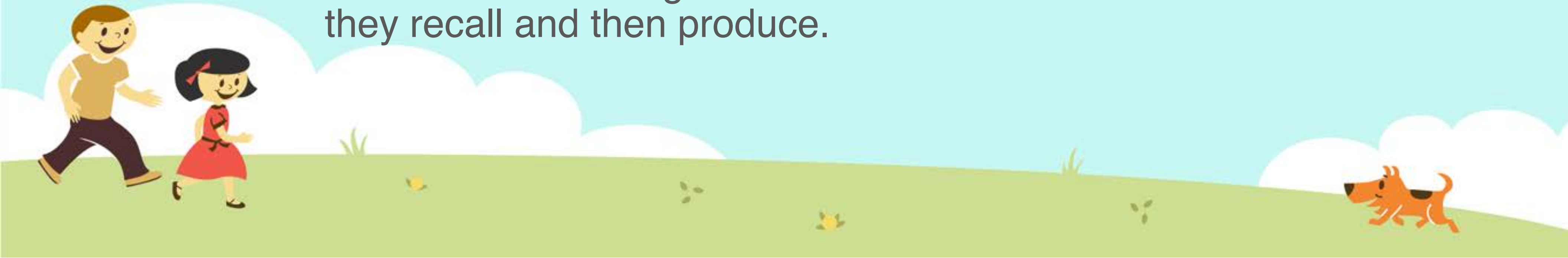
Presented by Alandi Bates

Does performance demonstrate competence?

- Language use is dependent upon maturational and experiential factors that impact linguistic representation. Therefore, it can be difficult to draw conclusions about how children accomplish the task of linguistic representation based solely upon what they say. This is demonstrated by Chomsky (1965) in the competence/ performance distinction.
- Shipley, Gleitman & Smith (1969) add support to this idea by showing that children in the so-called telegraphic stage understand fully and syntactically formed sentences better than ones that resemble their own productions.
  - If what children say reflects what they know then we would expect to see the opposite.

Do children have a productive linguistic system or do they use item-based schemas?

- The Continuity Hypothesis postulates that without evidence to the contrary it is to be assumed that children have fully productive grammars and linguistic representations that match that of adults. The divergence in production lies in the cognitive, perceptual, and articulatory limitations that constrain children (Yang 2010).
- The Item or Usage-Based Theory (Tomasello 1992) states that children's grammar is not productive and they do not share the same linguistic representations that adults have. Instead, they are believed to store linguistic forms as "chunks" of information that they recall and then produce.



Support for the Item or Usage-Based Theory

- Three case study examples given by Tomasello (2000) in support of the theory
  - The Verb Island Hypothesis
    - Early “syntactic competence is comprised totally of verb-specific constructions with open nominal slots”
  - Limited Morphological Inflection
    - 47% of verbs used in 1 person-number agreement where 6 forms are possible
  - Unbalanced Determiner Usage
    - When children first begin using the determiners a and the there is virtually no overlap in nouns and determiners used: if the cat is produced then a cat is not

Problems with the evidence for Item or Usage-Based Theory

- Based on “Intuitive Inspections” rather than formal empirical tests
- Observations haven’t been shown empirically to demonstrate that they are contradictory to a fully productive grammar
- Observations haven’t been shown empirically to accord with the theory itself
- This is perhaps because the theory isn’t defined clearly enough to enable a quantitative analysis of the results

Statistical properties of natural language
Zipfian Words

- Zipf’s Law (1949) hints that like many other things in the natural world language seems to follow an interesting trend, where a relatively small portion of all the words comprise an astonishingly large portion of the words produced and the remaining portion of words are produced infrequently and in many cases only once.
  - “More precisely, the frequency of a word tends to be approximately inversely proportional to its rank in frequency (Yang 2010)”

f = frequency of the word

r = rank of word in a sample of N words

$$f = \frac{C}{r} \quad \text{Where C is some constant} \qquad (1)$$

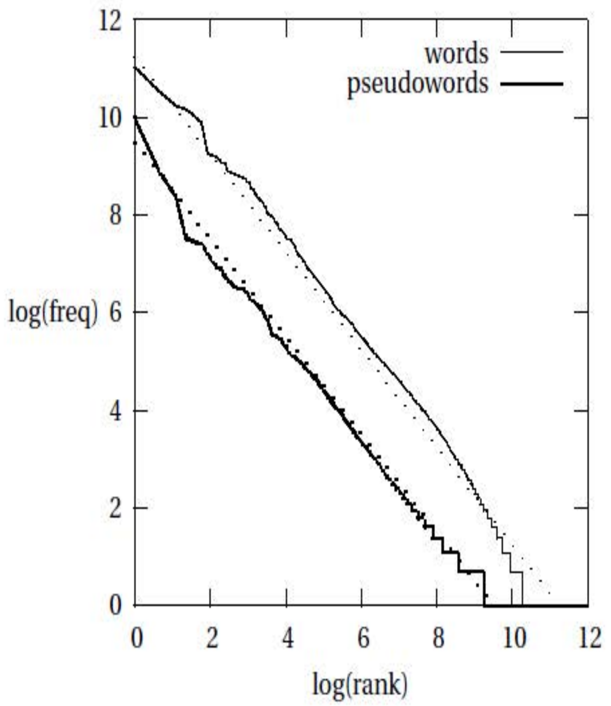


Figure 1. Zipfian distribution of words (top) and pseudowords (bottom) in the Brown corpus. The lower line is plotted by taking "words" to be any sequence of letters between *e*'s (Chomsky 1958). The two straight dotted lines are linear functions with the slope -1, which illustrate the goodness of the Zipfian fit.

Brown Corpus

Rank 1: the

$f \approx 70{,}000$

Rank 2: of

$f \approx 36{,}000$

* a word with rank 2 should appear ½ as often as the word with rank 1

$$f = \frac{C}{r} \qquad (1)$$

"By taking the log on both sides of the equation above

$(\log f = \log C - \log r)$

a perfect Zipfian fit would be a straight line with the slope -1 (Yang 2010)"

However, the actual frequency of the word doesn’t provide us with much information other than how often it occurs. What would be more informative as to whether the data support the idea that children have a productive grammar or an item-based schema of representation is knowing the probability of occurrence that each word has

The above equation (1) can be modified as such:

$$p_r = \left(\frac{C}{r}\right) / \left(\sum_{i=1}^{N} \frac{C}{i}\right) = \frac{1}{rH_N} \; \textit{where } \mathrm{H}_N \text{ is the Nth Harmonic Number } \sum_{i=1}^{N} \frac{1}{i} \qquad (2)$$

“Relatively little attention has been given to the combinatorics of linguistic units under a grammar and more important, how one might draw inference about the grammar given the distribution of word combinatorics. We turn to these questions immediately(Yang 2010)”

n-grams and syntactic rules of modern English: Zipfian Combinatorics

Words: 43% occur only 1x
58% occur 1-2 x's
68% occur 1-3 x's

The most frequent rule is
"PP→P NP"

Prepositional Phrase→Preposition Noun Phrase

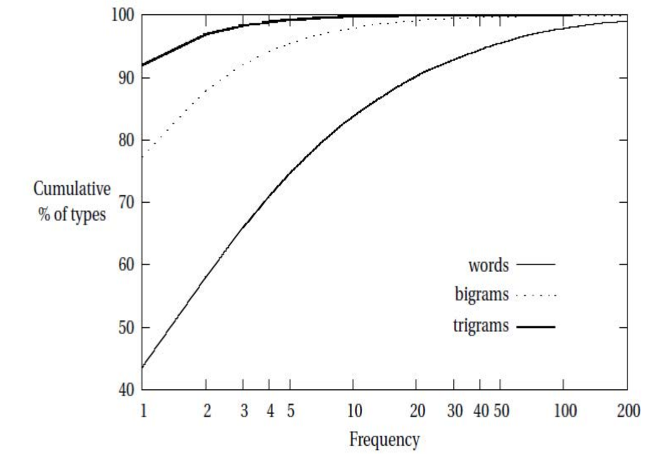


Figure 2. The vast majority of *n*-grams are rare events. The x-axis denotes the frequency of the gram, and the y-axis denotes the cumulative % of the gram that appear at that frequency or lower.

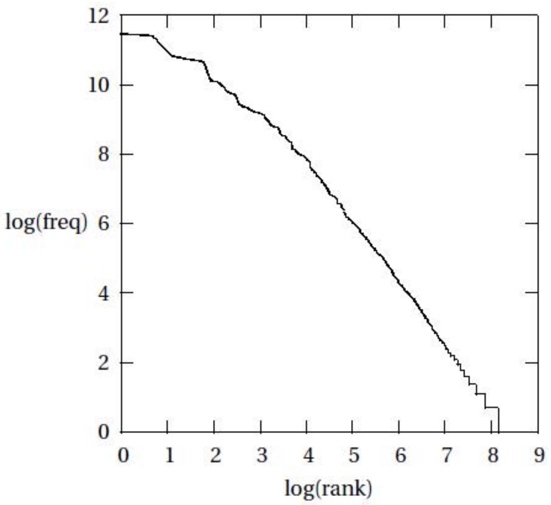


Figure 3. The frequency distribution of the syntactic rules in the Penn Treebank.

n-grams and syntactic rules of modern English: Zipfian Combinatorics

Bigrams: about 78% occur 1x
about 87% occur 1-2 x's
about 91% occur 1-3 x's

Followed by "S→NP VP"
Sentence→Noun Phrase Verb Phrase

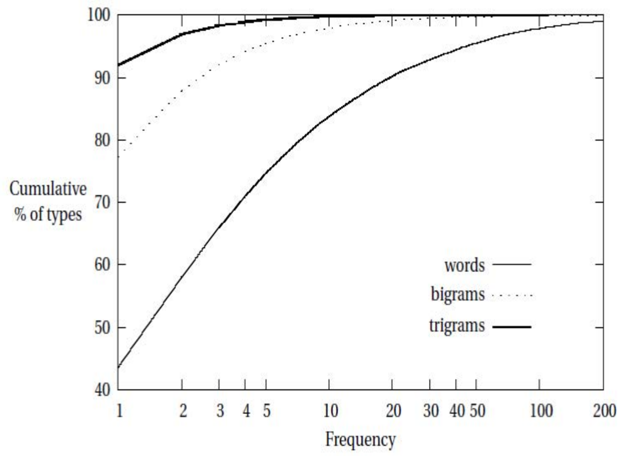


Figure 2. The vast majority of *n*-grams are rare events. The x-axis denotes the frequency of the gram, and the y-axis denotes the cumulative % of the gram that appear at that frequency or lower.

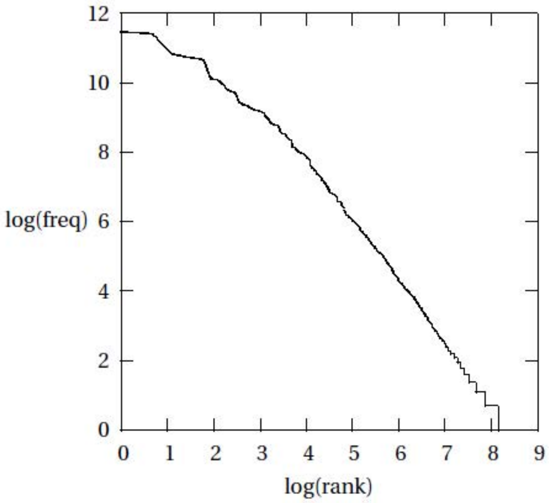


Figure 3. The frequency distribution of the syntactic rules in the Penn Treebank.

n-grams and syntactic rules of modern English: Zipfian Combinatorics

Trigrams: approx. 91% occur 1 x
approx. 96% occur 1-2 x's

“Certain rules have been collapsed together”

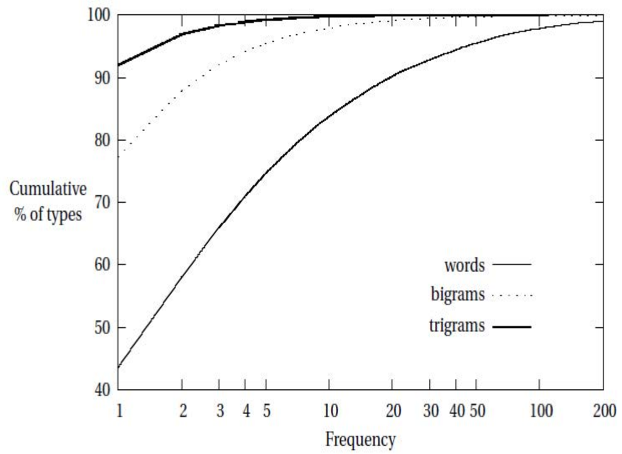


Figure 2. The vast majority of *n*-grams are rare events. The x-axis denotes the frequency of the gram, and the y-axis denotes the cumulative % of the gram that appear at that frequency or lower.

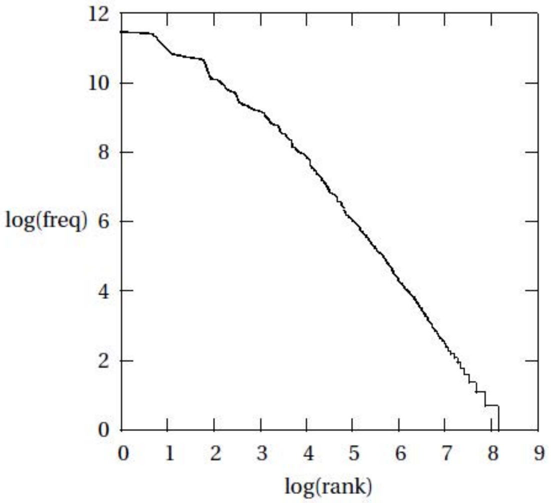
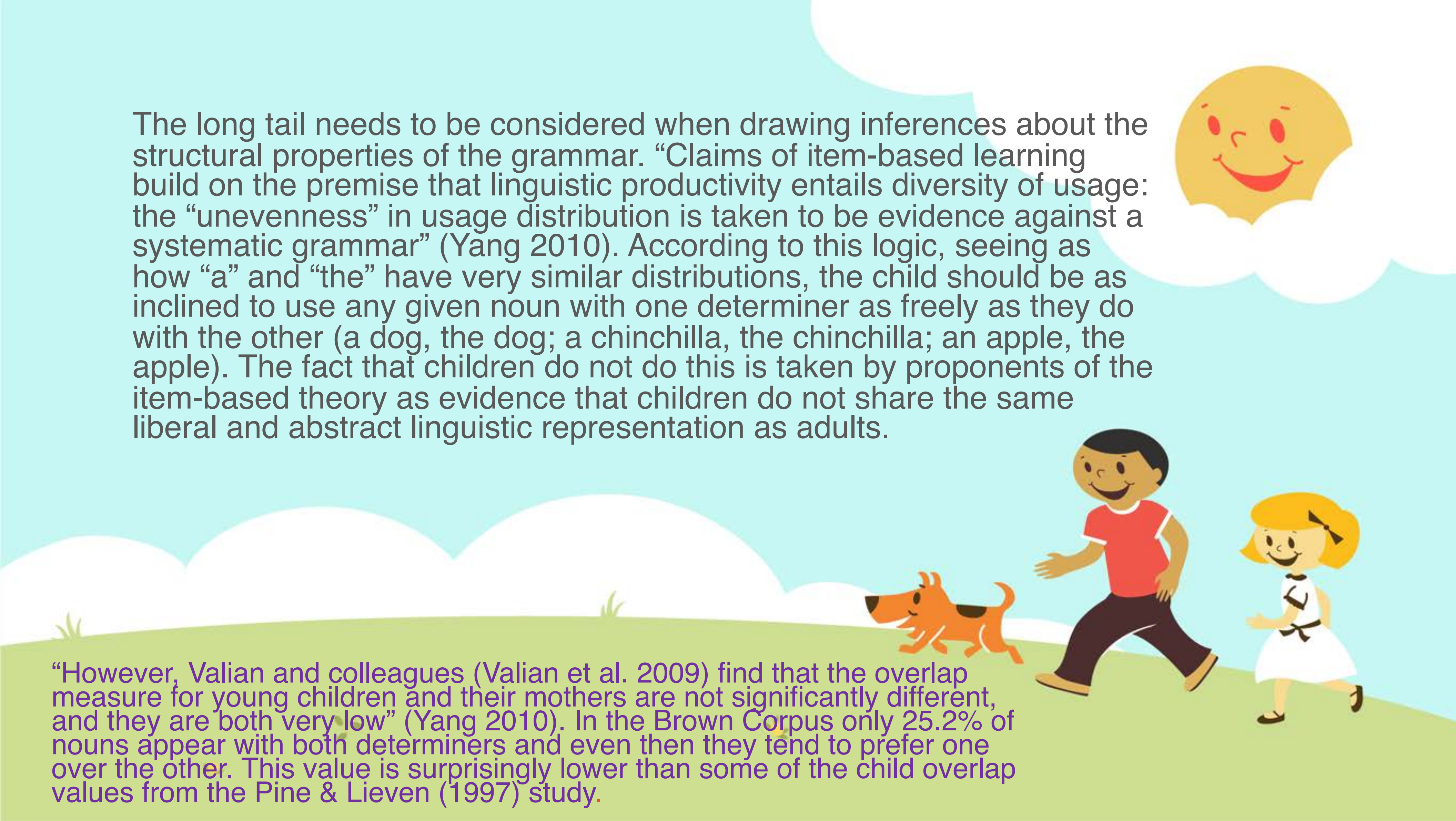


Figure 3. The frequency distribution of the syntactic rules in the Penn Treebank.

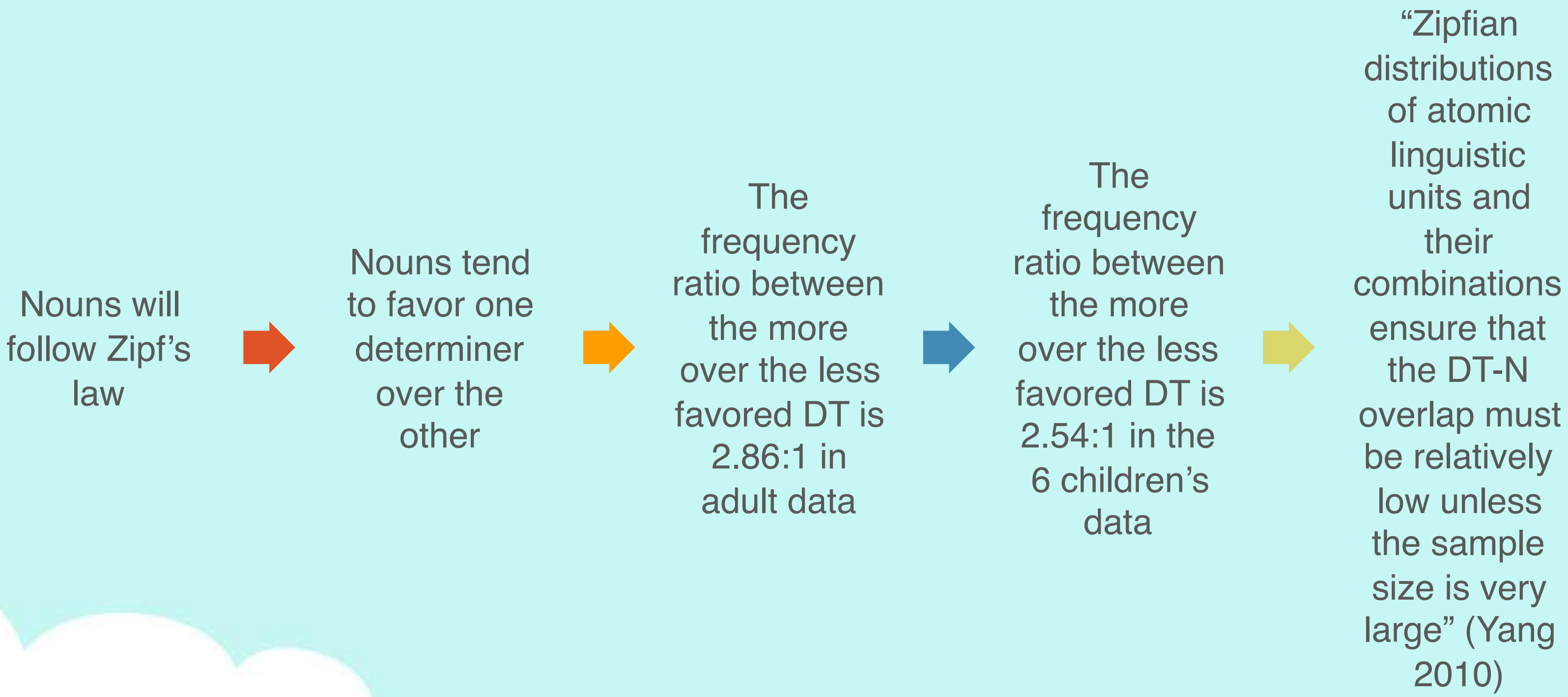


The long tail needs to be considered when drawing inferences about the structural properties of the grammar. “Claims of item-based learning build on the premise that linguistic productivity entails diversity of usage: the “unevenness” in usage distribution is taken to be evidence against a systematic grammar” (Yang 2010). According to this logic, seeing as how “a” and “the” have very similar distributions, the child should be as inclined to use any given noun with one determiner as freely as they do with the other (a dog, the dog; a chinchilla, the chinchilla; an apple, the apple). The fact that children do not do this is taken by proponents of the item-based theory as evidence that children do not share the same liberal and abstract linguistic representation as adults.

“However, Valian and colleagues (Valian et al. 2009) find that the overlap measure for young children and their mothers are not significantly different, and they are both very low” (Yang 2010). In the Brown Corpus only 25.2% of nouns appear with both determiners and even then they tend to prefer one over the other. This value is surprisingly lower than some of the child overlap values from the Pine & Lieven (1997) study.

- This suggests that not only is adult productivity poor it is more demonstrative of item-based learning than small children. How can children, who according to Tomasello, Pine & Lieven, and others, don’t have adult like grammar representations, appear to use the DT category more productively than adults?
  - It’s because of the “Zipfian distribution of syntactic categories and the generative capacity of natural language grammar” (Yang 2010).
    - DP→DT N
      - In accordance with productivity the two components combine independently of each other.
      - Few nouns appear often while some appear only once thereby decreasing the opportunity to appear with more than 1 DT; no overlap.

So what can we conclude about the distributions of the 2 categories and their combinations?



Nouns will follow Zipf's law

→

Nouns tend to favor one determiner over the other

→

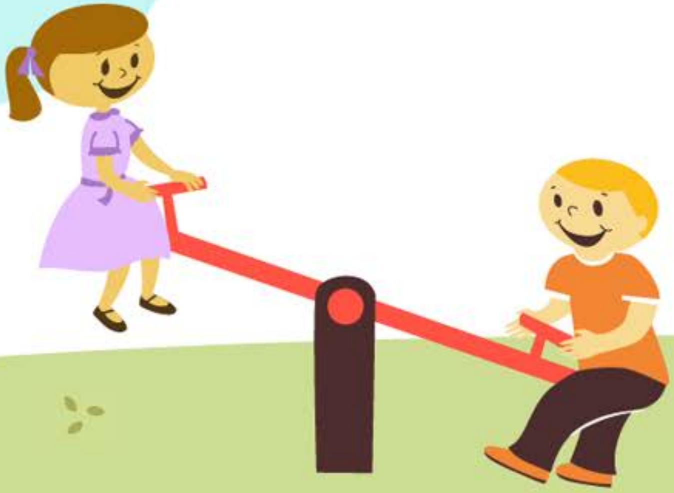
The frequency ratio between the more over the less favored DT is 2.86:1 in adult data

→

The frequency ratio between the more over the less favored DT is 2.54:1 in the 6 children's data

→

"Zipfian distributions of atomic linguistic units and their combinations ensure that the DT-N overlap must be relatively low unless the sample size is very large" (Yang 2010)



Quantifying Productivity

Analysis of DT N overlap values

Under the productive rule DP→DT N

n_r following equation (2) has the Zipfian probability

$p_r = \dfrac{1}{rH_N}$ of being drawn at any single trial in S.

expected overlap value of

n_r **O(r, N, D, S)**

Overlap for the sample

$$O(D,N,S) = \frac{1}{N}\sum_{i=1}^{N} O(r,N,D,S) \qquad (3)$$

Expected value of the overlap value for the sample (N,D,S) under the productive rule “DP→D N

"Consider a sample (N,D,S), which consists of N unique nouns, D unique determiners, and S determiner-noun pairs. D=2 The nouns that have appeared with more than one will have an overlap value of 1; otherwise, they have an overlap value of 0. The overlap value for the entire sample will be the # of 1's divided by N" (Yang 2010).

Consider now the calculation O(r,N,D,S) Since n_r has an overlap value of 1 if and only if
It has been used with more than one determiner in the sample

$$O(r,N,D,S) = 1 - \Pr \ \{\text{where } n_r \text{ is not sampled during S trials}\}$$

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

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

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

In the above portion of the equation the determiner and noun are independent. Therefore, the probability of noun n_r combining with the i th determiner is the product of their probabilities $d_i p_r$

Which is represented by the multinomial expression below

$$(p_1 + p_2 + \ldots + p_{r-1} + d_i p_r + p_{r+1} + \ldots + p_N)^S$$

However, this value includes the probability of the word combining with the i th determiner zero times $(1 - p_r)^S$ and therefore needs to be subtracted off

Thus the probability with which n_r combines with the i th determiner exclusively in the sample S is: $[(d_i p_r + 1 - p_r)^S - (1 - p_r)^S]$

Summing these values over all determiners and collecting terms, we have:

$$O(r,N,D,S) = 1 + (D-1)(1-p_r)^S - \sum_{i=1}^{D}[(d_i p_r + 1 - p_r)^S]$$

This allows for the calculation of the expected value of overlap using only the sample size S, the number of unique noun N and the number of unique determiners D, under the assumption that nouns and determiners both follow Zipf’s law

What amount of overlap should we expect to find?

- Low overlap values are mathematically necessary according to the distribution.

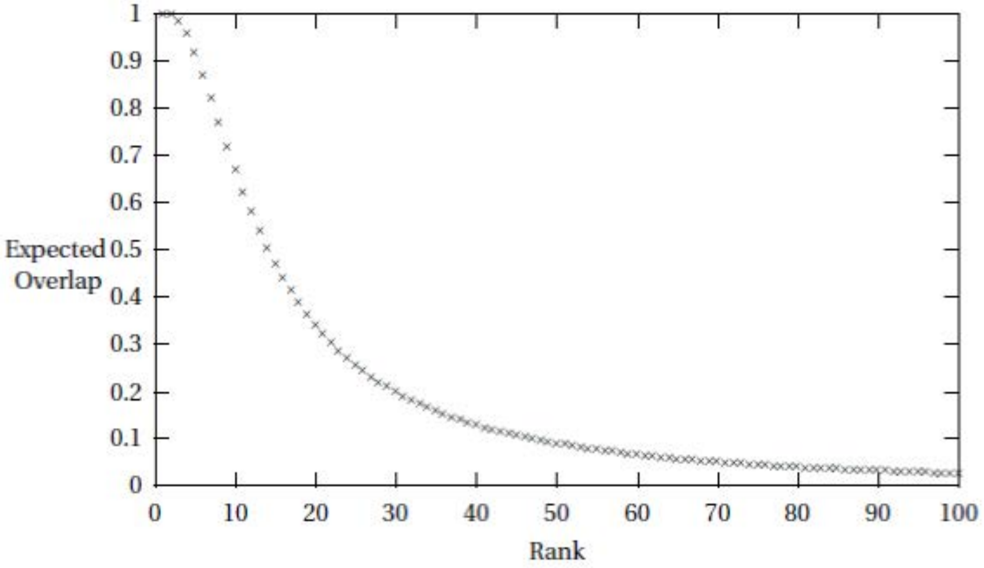
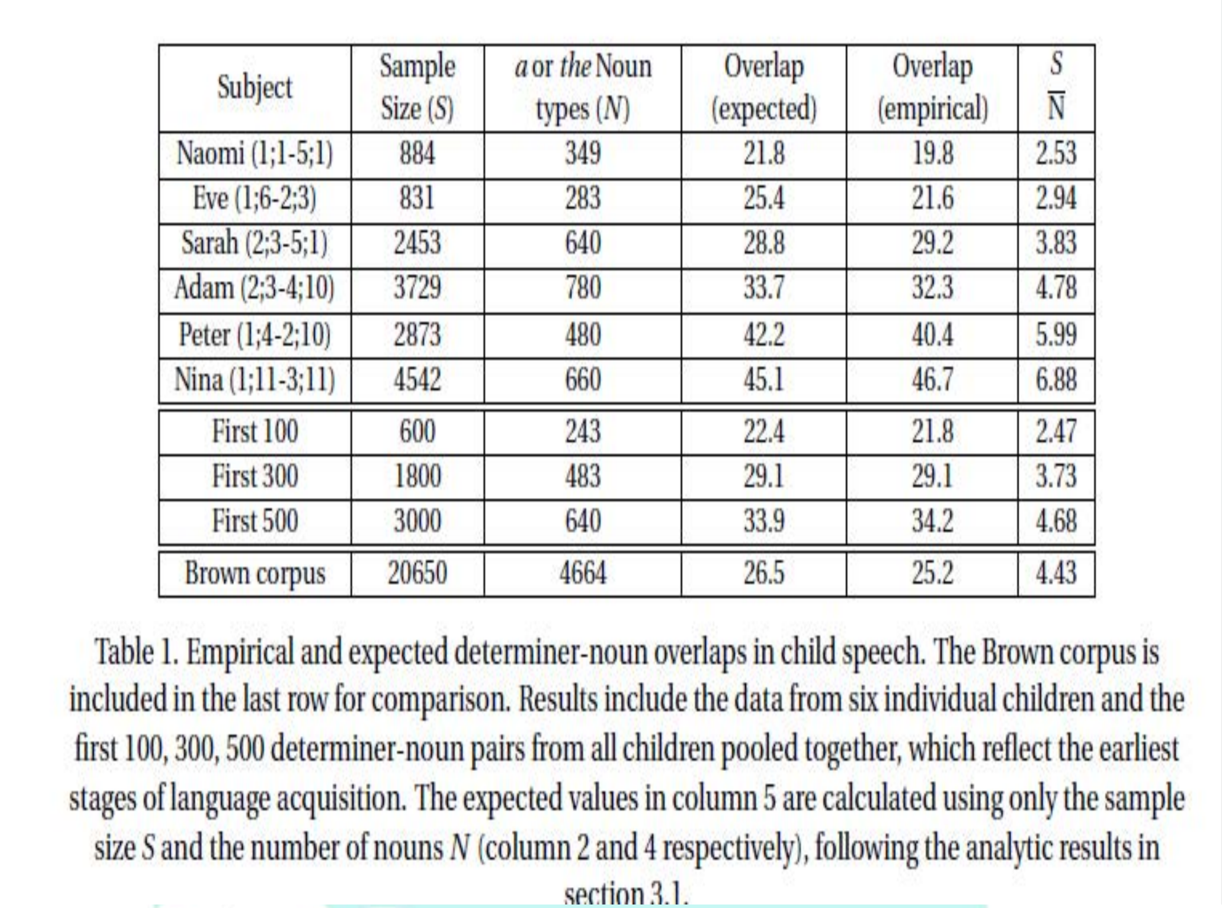


Figure 4. Expected overlap values for nouns ordered by rank, for $N = 100$ nouns in a sample size of $S = 200$ with $D = 2$ determiners. Word frequencies are assumed to follow the Zipfian distribution. As can be seen, few of nouns have high probabilities of occurring with both determiners, but most are (far) below chance. The average overlap is 21.1%.

Determiners and Productivity: Methods

- Used data from 6 children in the CHILDES database: Adam, Eve, Sarah, Naomi, Nina, and Peter
- These were the only children in the database with substantial longitudinal data of the early stages of syntactic development so that the item-based stage, if exists, could be observed.
- For comparison the Brown Corpus (Kucera & Francis 1967) and its overlap measures were used
- As per a suggestion by Virginia Valian, all the child data for the first 100, 300, and 500 were pooled together to create 3 hypothetical children that represented three stages of syntactic development
- The results are summarized in the table on the next slide



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

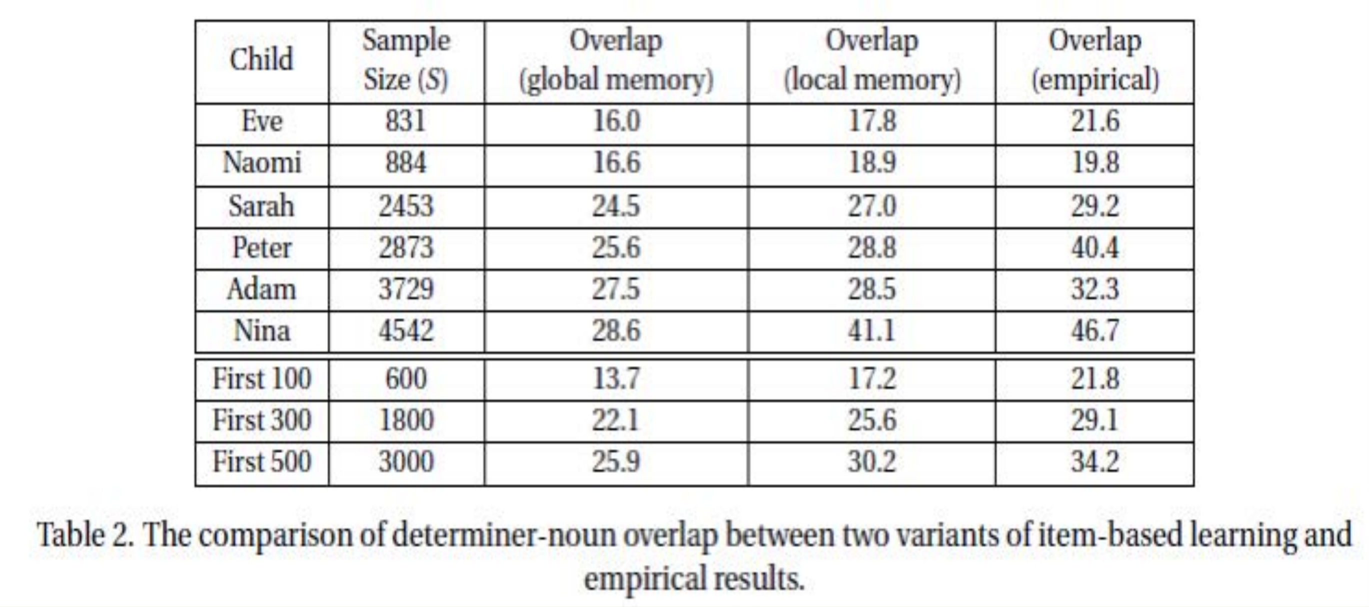
Table 1. Empirical and expected determiner-noun overlaps in child speech. The Brown corpus is included in the last row for comparison. Results include the data from six individual children and the first 100, 300, 500 determiner-noun pairs from all children pooled together, which reflect the earliest stages of language acquisition. The expected values in column 5 are calculated using only the sample size S and the number of nouns N (column 2 and 4 respectively), following the analytic results in section 3.1.

Evaluating Item-Based learning

- Because of the lack of concrete models it's hard to evaluate the item-based learning theory. However, a plausible approach can be used to approximate the central tenet of the theory that children "memorize and retrieve specific itemized combinations.
- They consider a learning model that memorizes jointly formed, as opposed to productively composed, D-N pairs from the input.
- There are 2 models:
  - Global Memory Model: a hypothetical child that memorizes the DN combos and their frequencies from the 1.1 million random sample of English adult utterances from the CHILDES database
  - Local Memory Model: a hypothetical child that memorizes the DN combos and their frequencies from a particular child's CHILDES transcript of adult utterances
  - The Monte Carlo simulation was used with each child and each variant of the memory model to randomly draw S pairs from the 2 sets of data that correspond to the local and global memory learning models
  - The more frequently a pair occurs the greater probability of it being drawn, which aligns with the frequency effects of the item/usage-based approach
  - Results are averaged over 1000 draws

Both sets of overlap values from the 2 variants of item-based learning differ significantly from the empirical measures

- Children’s use of determiners does not follow the item-based learning theory, at least as far the models were used to exploit the memorization of jointly formed DN pairs. Since the theoretical predictions of the theory aren’t well defined these results are tentative until such time when the appropriate test can be utilized to evaluate the theory



Child	Sample Size (*S*)	Overlap (global memory)	Overlap (local memory)	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

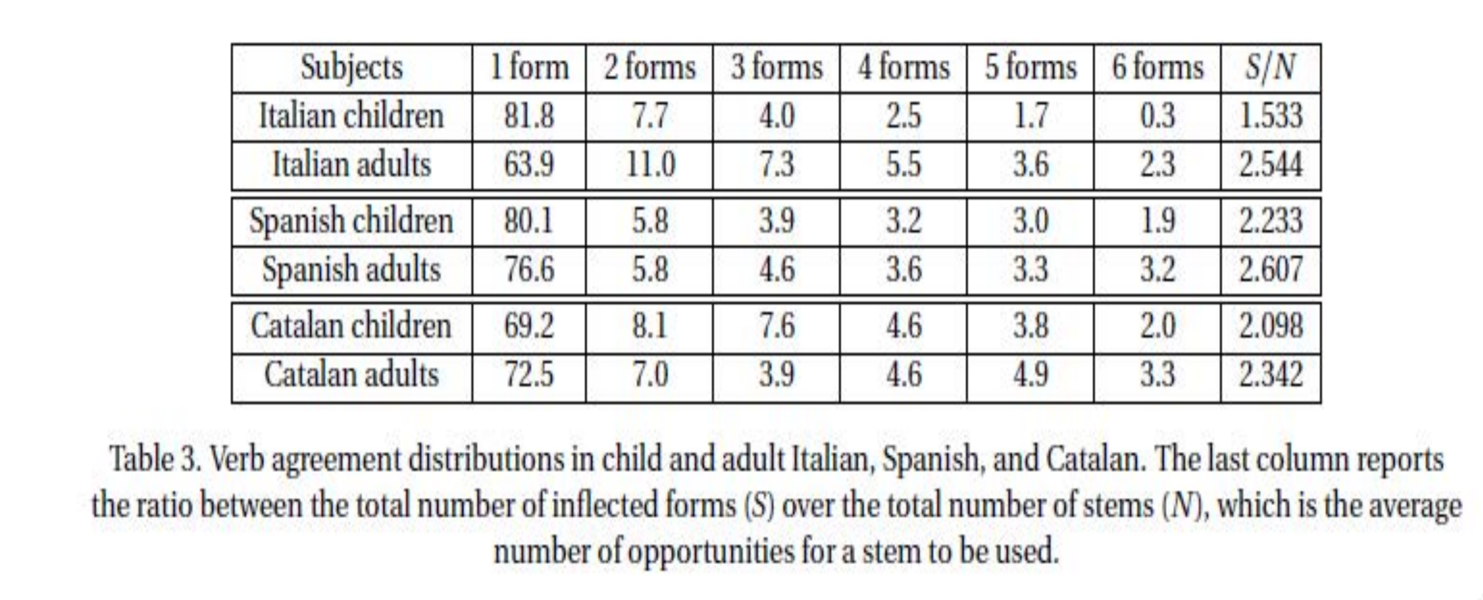
Table 2. The comparison of determiner-noun overlap between two variants of item-based learning and empirical results.

An Itemized look at VERBS

- “The Zipfian reality is inherent : the combinatorics of verbs and their morphological and syntactic associates are similarly lopsided in usage distribution as with the determiners”
- “Few stems appear in a great number of inflections, which, never approach anywhere near the maximum number of possible inflections..most stems are used sparsely, the majority of which occur in exactly 1 inflection”.
- “Furthermore, the inflections themselves are also Zipfian: few are used very frequently but most are used sparsely”

The diversity of usage depends on the # of opportunities for a verb stem to appear in multiple forms

- “Each cell represents the percentage of verb stems that are used in 1,2,3,4,5, and 6 Inflectional forms

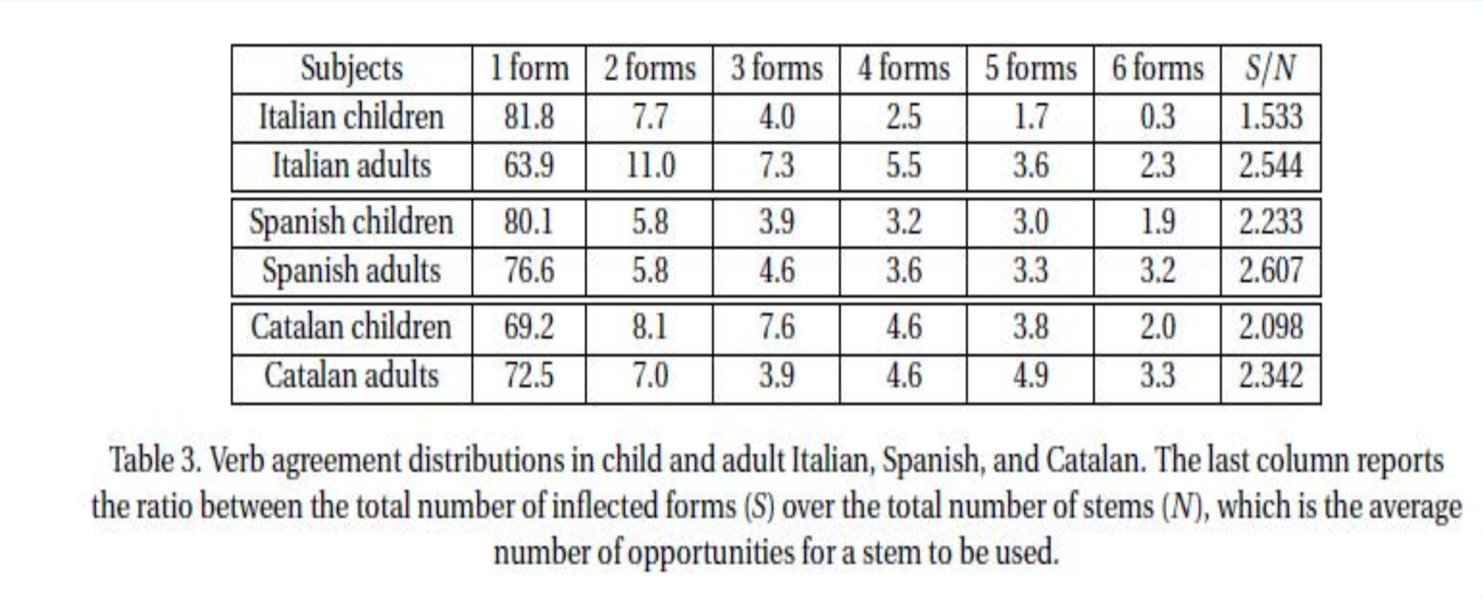


Subjects	1 form	2 forms	3 forms	4 forms	5 forms	6 forms	S/N
Italian children	81.8	7.7	4.0	2.5	1.7	0.3	1.533
Italian adults	63.9	11.0	7.3	5.5	3.6	2.3	2.544
Spanish children	80.1	5.8	3.9	3.2	3.0	1.9	2.233
Spanish adults	76.6	5.8	4.6	3.6	3.3	3.2	2.607
Catalan children	69.2	8.1	7.6	4.6	3.8	2.0	2.098
Catalan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

Table 3. Verb agreement distributions in child and adult Italian, Spanish, and Catalan. The last column reports the ratio between the total number of inflected forms (S) over the total number of stems (N), which is the average number of opportunities for a stem to be used.

Spanish and Catalan learning children show similar agreement usage to that of adults

- “Each cell represents the percentage of verb stems that are used in 1,2,3,4,5, and 6 Inflectional forms

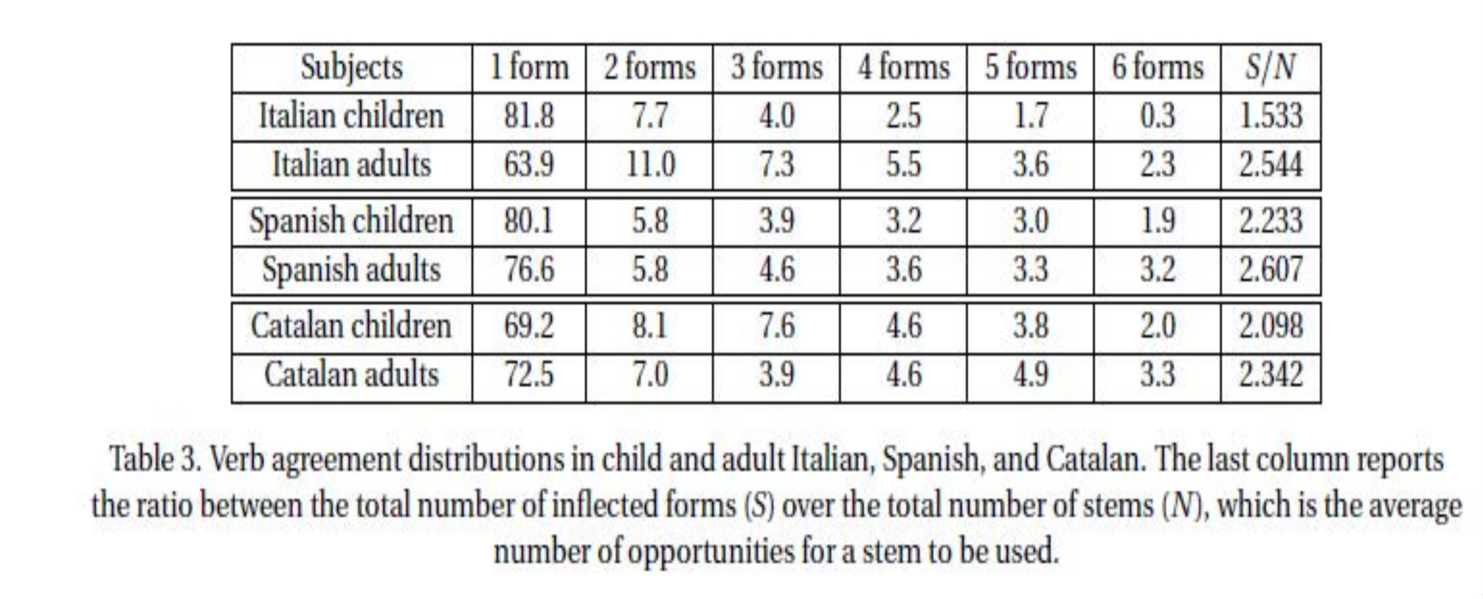


Subjects	1 form	2 forms	3 forms	4 forms	5 forms	6 forms	S/N
Italian children	81.8	7.7	4.0	2.5	1.7	0.3	1.533
Italian adults	63.9	11.0	7.3	5.5	3.6	2.3	2.544
Spanish children	80.1	5.8	3.9	3.2	3.0	1.9	2.233
Spanish adults	76.6	5.8	4.6	3.6	3.3	3.2	2.607
Catalan children	69.2	8.1	7.6	4.6	3.8	2.0	2.098
Catalan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

Table 3. Verb agreement distributions in child and adult Italian, Spanish, and Catalan. The last column reports the ratio between the total number of inflected forms (S) over the total number of stems (N), which is the average number of opportunities for a stem to be used.

Italian children use more stems in only one inflection than the adults but this follows from the S/N ratio (2.544 vs. 1.533). The adults have roughly 66% more opportunities to use it than the children which accounts for the discrepancy

- “Each cell represents the percentage of verb stems that are used in 1,2,3,4,5, and 6 Inflectional forms



Subjects	1 form	2 forms	3 forms	4 forms	5 forms	6 forms	S/N
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

Table 3. Verb agreement distributions in child and adult Italian, Spanish, and Catalan. The last column reports the ratio between the total number of inflected forms (S) over the total number of stems (N), which is the average number of opportunities for a stem to be used.

All verbs are islands

- “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 (1992, p242), each unique lexical item in the object position counts as a unique construction for the verb”
  - Top 15 transitive verbs in 1.1 million child directed utterances

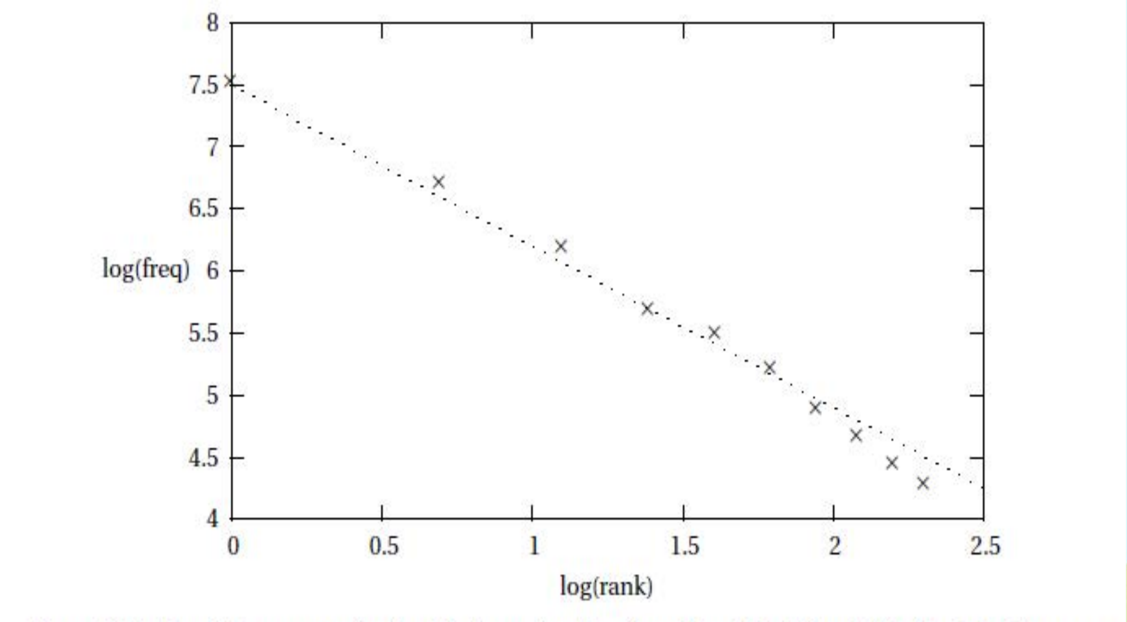


Figure 6. Rank and frequency of verb-object constructions based on 1.1 million child-directed utterances.

“even for large corpora, a verb appears in few constructions frequently and in most constructions infrequently if at all. The observation of Verb Islands, is in fact characteristic of a fully productive verbal syntax system”

All verbs are islands

- “The quantitative predictions of the Verb Island Hypothesis have not been spelled out but we may estimate the necessary amount of language sample that would mask these effects”
- “Instead of calculating the expected numbers of determiners that a noun appears with, one would calculate the expected number of objects a verb appears with”.
- Monte Carlo simulation to determine range of sample size
  - For 100 verbs and 100 objects sample size would have to be approx. 28,000 verb-object pairs to meet 50% requirement
  - For 1500 verbs and 1500 objects sample size would be 4.8 billion words to meet 50% requirement, which would amount to 46 years of non-stop talking at a rate of 200 wds/min
  - How can you find anything but verb islands

What we can conclude from the information discussed and presented

- “While the current study show that children’s production is consistent with a productive grammar, it should not distract us from the important question how the child learns that grammar in the first place”
- “For the linguist, the Zipfian nature of language raises important questions for the development of linguistic theories”
  - 1st: “Zipf’s law hints at the inherent limitations in approaches that stress the storage of a construction-specific rules orprocesses”
  - 2nd: “Zipf’s law challenges the conventional wisdom in current syntactic theorizing that makes use of a highly detailed lexical component

Final Thought

“The most significant victim of George Kingsley Zipf, must be the child herself. The task faced by children acquiring language is no different than from that of the psychologist, computer scientist, and linguist, for the input data are also Zipfian in character, except the child does not have the tools that the above mentioned have. The sparse data problem strikes just as hard, and the role of memory in language learning should not be overestimated. The learner’s Zipfian challenge bears another name: the argument from the POVERTY OF THE STIMULUS. In the face of such statistical reality of language, a grammatical system with full generative potentials from the get go still seems the best preparation a child can hope for” (Yang 2010)