Supporting Information

Yang 10.1073/pnas.1216803110

SI Text

Data and Empirical Methods

The child language data are drawn from the publicly available CHILDES database (1). Six children ("Adam," "Eve," "Naomi," "Nina," "Peter," and "Sarah") are selected; these are the only American English-learning children in the public domain whose data start at the very beginning of syntactic combinations and make up sufficiently large samples for statistical analysis. The age spans of these children are given in Table S1. The data are processed using a state-of-the art part-of-speech (POS) tagger (http://gposttl.sourceforge.net). The Brown Corpus is publicly available with words already annotated with POSs. All determiner-noun pairs are extracted from the tagged word sequences in which the first word is a determiner ("a" or "the") and the second word is tagged as a singular noun. In addition, I pooled the first 100, 300, and 500 determiner-noun pairs produced by each child to create three composite learners, which collectively represent the very earliest data on child language in the public domain. The ages (year and month) of the children at these cutoff points are: Adam (2;6, 2;9, 3;0), Eve (1;8, 1;11, 2;1), Naomi (2;0, 2;7, 2;11), Nina (1;11, 2;0, 2;1), Sarah (2;8, 3;1, 3;3), and Peter (2;0, 2;1, 2;2). These children are clearly at variant stages of language development, all of which are well explained by the statistical model developed here. Nim's sign combination data are taken from work by Terrace (2).

Zifp's Law and Language

Under a perfect fit of Zipf's law, word ranks and frequencies have the slope of −1.0 on the log-log scale. Studies across languages and genres strongly confirmed the accuracy and universality of Zipf's law (3). For the child language data used here, the average slope of the linear fit is −0.98, again pointing to the accuracy of Zipf's law. Thus, I can approximate the marginal probabilities of words using Eq. **1**: that a word with a rank r in a sample of N words has a probability of $1/(rH_N)$, where H_N is the harmonic number $\sum_{i=1}^{N} 1/i$.

As noted in the main text, only 25.2% of singular nouns in the Brown Corpus appear with both "a" and "the." Similar patterns hold for children's speech data. On average, 22.8% of the nouns in each sample appear with both "a" and "the." For these, the more vs. less favored determiner has an average frequency ratio of 2.54:1. The identity of the favored determiner varies from noun to noun, as the example of "bath/bathroom" from the main text makes clear.

These results suggest that Zipf's law characterizes the frequencies of words as the propensities of word combinations.

Statistics of Grammar

The calculation uses the determiner-noun example but is applicable to any combinations of linguistic units. A productive rule "NP→DN," where NP is a noun phrase, D is a determiner, and N is a noun, means that the combination of categories (determiner and noun) is independent. Let the marginal probability of drawing the noun n_r, $1 \le r \le N$ in each trial be p_r, and let that of drawing the ith determiner be d_i. The expected probability of n_r being drawn with both determiners, E_r, is as follows:

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

The last term above requires a brief comment. The independence of determiner-noun combinations under the rule means that the probability of the noun n_r following the ith determiner is the product of their probabilities, or $d_i p_r$. The multinomial expression

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

gives the probabilities of all the compositions of the sample, with n_r combining with the ith determiner 0, 1, 2, … S times, which is $(d_i p_r + 1 - p_r)^S$ because $(p_1 + p_2 + \ldots + p_{r-1} + p_r + p_{r+1} + \ldots + p_N) = 1$. However, this value includes the probability of n_r combining with the ith determiner zero times; again $(1 - p_r)^S$, which must be subtracted. Thus, the probability with which n_r combines with the ith determiner exclusively in the sample S is $[(d_i p_r + 1 - p_r)^S - (1 - p_r)^S]$.

The average of the entire sample is Eq. **2**, repeated below:

$$E[D] = \frac{1}{N} \sum_{r=1}^{N} E_r$$

Child Language, Grammar, and the Memory Model of Language Learning

The data and the syntactic diversity measures are provided in Table S1. The results from the memory-based learning model, which implements a suggestion in the study by Tomasello (4), are included there also.

High diversity of determiner-noun combinations can only be obtained when more nouns are sampled more than once so that they may have an opportunity to combine with multiple determiners. If noun probabilities follow Zipf's law (3), S/H_N nouns are expected to occur more than once, and the average diversity over all N nouns is thus positively correlated with $S/(NH_N)$, which is approximately $S/(N \ln N)$ because $H_N \approx \ln N$ [a similar analysis is given in a study by Valian et al. (5)]. Given the slow asymptotic growth of ln N, the ratio S/N (average number of times a noun is used in the speech sample) predicts the diversity measure. This is strongly confirmed ($\rho = 0.985$, $P < 10^{-5}$) for the data in Table S1.

Nim's Sign Combinations

Nim's two sign combinations are grouped into eight potential rules (2). Each rule consists of a closed class functor (one of two words, such as "more" and "give") followed or preceded by an open class item, generating patterns such as "more apple," "give apple," and "more ball." The theoretical value of diversity is calculated assuming the combination of the two categories in the rule is independent.

An alternative calculation ignores the word order restrictions in Nim's constructions; for instance, an open class item "banana" is considered to have been paired with "more" and "give" regardless of their relative positions. There are now four instead of eight constructions. Doing so increases the sample size for each construction, but the increase in the types of open class items is very modest. Consequently, Nim's combinatorial diversities for constructions without word order are 46.5%, 68.6%, 94%, and 60%, respectively, whereas the expected values are 90.7%, 99.9%, 99.9%, and 80.9%, respectively. It seems clear that Nim's productivity still falls far short of what could be expected of a productive combinatorial system, even if I relax the restrictions on word order.

The disproportionally large sample size over few types accounts for Nim's much higher diversity values than those of human subjects. If Nim imitated his teachers' sign combinations as suggested by Terrace et al. (6), Nim had ample opportunities to copy from his (productive) sign language teachers.

1. MacWhinney B (2000) *The CHILDES Project* (Lawrence Erlbaum, Mahwah, NJ).
2. Terrace HS (1979) *Nim* (Knopf, New York).
3. Baroni M (2009) *Corpus Linguistics*, eds Lüdelign A, Kytö M (Mouton de Gruyter, Berlin), pp 803–821.
4. Tomasello M (2000) First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11(1–2):61–82.
5. Valian V (1986) Syntactic categories in the speech of young children. *Dev Psychol* 22: 562–579.
6. Terrace HS, Petitto LA, Sanders RJ, Bever TG (1979) Can an ape create a sentence? *Science* 206(4421):891–902.

Table S1. Empirical values of syntactic diversity in human language compared with theoretical values of a productive grammar model and a memory-based learning model

Subject	Sample size	Types	Theoretical, %	Empirical, %	Memory model, %
Naomi (1;1–5;1)	884	349	21.8	19.8	16.6
Eve (1;6–2;3)	831	283	25.4	21.6	16.0
Sarah (2;3–5;1)	2,453	640	28.8	29.2	24.5
Adam (2;3–4;10)	3,729	780	33.7	32.3	27.5
Peter (1;4–2;10)	2,873	480	42.2	40.4	25.6
Nina (1;11–3;11)	4,542	660	45.1	46.7	28.6
First 100	600	243	22.4	21.8	13.7
First 300	1,800	483	29.1	29.1	22.1
First 500	3,000	640	33.9	34.2	25.9
Brown Corpus	20,650	4,664	26.5	25.2	n/a

The notation X;Y indicates the child's age (year and month). The Brown Corpus is also examined for comparison. The agreement between the theoretical and empirical values (columns 4 and 5) is statistically significant (concordance correlation coefficient $\rho_c = 0.977$, 95% confidence interval: 0.925–0.993). The final column shows the simulation results (averaged over 1,000 trials) of the memory model described in *SI Text*, which are significantly lower than the empirical values ($P < 0.002$, paired one-tailed Mann–Whitney test). n/a, not applicable.

Table S2. Empirical values of syntactic diversity in Nim's eight sign combination patterns compared with theoretical values if the combinations are independent

Rule	Sample size	Types	Theoretical, %	Empirical, %
(more|give) X	1,215	67	88.0	44.8
X (more|give)	256	39	59.9	35.9
Verb (me|Nim)	800	14	99.9	78.6
(me|Nim) verb	158	13	87.4	46.1
Food-item (me|Nim)	775	18	99.7	88.9
(me|Nim) food-item	261	14	94.9	85.7
Nonfood-item (me|Nim)	180	20	75.9	70.0
(me|Nim) nonfood-item	99	19	57.2	36.8

The empirical values are significantly lower ($P < 0.004$, paired one-tailed Mann–Whitney test).