Ontogeny and phylogeny of language

Charles Yang[1]

Department of Linguistics and Computer Science, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19081

Edited* by William Labov, University of Pennsylvania, Philadelphia, PA, and approved February 14, 2013 (received for review September 26, 2012)

How did language evolve? A popular approach points to the similarities between the ontogeny and phylogeny of language. Young children's language and nonhuman primates' signing both appear formulaic with limited syntactic combinations, thereby suggesting a degree of continuity in their cognitive abilities. To evaluate the validity of this approach, as well as to develop a quantitative benchmark to assess children's language development, I propose a formal analysis that characterizes the statistical profile of grammatical rules. I show that very young children's language is consistent with a productive grammar rather than memorization of specific word combinations from caregivers' speech. Furthermore, I provide a statistically rigorous demonstration that the sign use of Nim Chimpsky, the chimpanzee who was taught American Sign Language, does not show the expected productivity of a rule-based grammar. Implications for theories of language acquisition and evolution are discussed.

computational linguistics | linguistics | primate cognition | psychology

The hallmark of human language, and *Homo sapiens*' great leap forward, is the combinatorial use of language to create an unbounded number of meaningful expressions (1). How did this ability evolve? For a cognitive trait that inconveniently left no fossils behind, a popular approach points to the continuity between the ontogeny and phylogeny of language (2–4): "the most promising guide to what happened in language evolution," according to a comprehensive recent survey (5). Young children's language is similar to the signing patterns of primates: Both seem to result from imitation because they show limited and formulaic combinatorial flexibility (6, 7). Only human children will go on to acquire language, but establishing a common starting point before full-blown linguistic ability may reveal the transient stages in the evolution of language.

The ontogeny and phylogeny argument has force only if the parallels between primate and child language are genuine. Indeed, the assessment of linguistic capability, in both children and primates, has been controversial. Traditionally, children's language is believed to include abstract linguistic representations and processes even though their speech output may be constrained by nonlinguistic factors, such as working memory limitations. The relatively low rate of errors in many (but not all) aspects of child language is often cited to support this interpretation (8). A recent alternative approach emphasizes the memorization of specific strings of words rather than systematic rules (6, 9); the rarity of errors in child language could be attributed to memorization and retrieval of specific linguistic expressions in the adult input (which would be largely error-free).

The main evidence for learning by memorization comes from the relatively low degree of combinatorial diversity, which can be quantified as the ratio of attested vs. possible syntactic combinations. For instance, English singular nouns can interchangeably follow the singular determiners "a" and "the" (e.g., "a/the car," "a/the story"). If every noun that follows "a" also follows "the" in some sample of language, the diversity measure will be 100%. If nouns appear with "a" or "the" exclusively, the diversity measure will be 0%. Even at the earliest stage of language learning, children very rarely make mistakes in the use of determiner-noun combinations: Ungrammatical combinations (e.g., "the a dog," "cat the") are virtually nonexistent (10). However, the syntactic diversity of determiner-noun combinations is quite low: Only 20–40% of singular nouns in child speech appear with both determiners, and the rest appear with one determiner exclusively (11). These low measures of diversity have been interpreted as the absence of a systematic grammar: If the combination of determiners and nouns is truly independent and productive, a higher proportion of nouns may be expected to pair with both suitable determiners. However, subsequent studies show comparably low diversity measures in the speech of mothers, whose linguistic productivity is not in doubt (12). Perhaps more paradoxically, analysis of the Brown Corpus (13), a collection of English print materials, shows that only 25% of single nouns appear with both determiners, fewer than the diversity measure of very young children (11); it seems absurd to suggest that professional writers have a less systematic grammar than 2-y-old children.

The assessment of nonhuman primates' ability to learn language has also been riddled with controversies. Many earlier studies were complicated by researchers' subjective interpretations of behavioral data (reviewed in ref. 14). Project Nim was a notable exception (7). Nim Chimpsky was a chimpanzee taught American Sign Language (ASL) by human surrogate parents and teachers. Importantly, Nim's sign production data remain the only publicly available corpus from primate language research (15). Nim produced numerous sign combinations that initially appeared to follow a grammar-like system. However, further video analysis showed evidence of imitation of the teachers, leading the researchers to a negative assessment of Nim's linguistic ability (7). Yet video analysis of human primate interaction also contains an element of subjectivity, and the debate over primates' linguistic abilities continues (16).

These conflicting and paradoxical interpretations of child and primate languages are due, in part, to the absence of a statistically rigorous analysis of language use. What is the statistical profile of language if it follows grammatical rules? How is that distinguished from the statistical profile of language use by imitation? This paper develops a statistical test that can detect the presence or absence of grammatical rules based on a linguistic sample. I use this test to show that very young children's language is consistent with a grammar that independently combines linguistic units and is inconsistent with patterns of memorization of caregivers' speech. Furthermore, I show that Nim's sign combinations fall below the diversity expected of a rule-based grammar. I start with some statistical observations about language.

Statistics of Grammar

Zipf's Law and Language. It is well known that Zipf's law accurately characterizes the distribution of word frequencies (17). Let n_r be the word of rank r among N distinct words. Its probability,

Author contributions: C.Y. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]E-mail: charles.yang@ling.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1216803110/-/DCSupplemental.

P_r, of appearing in a corpus of N word types can be accurately approximated:

$$p_r = \left(\frac{C}{r}\right) \Big/ \left(\sum_{i=1}^{N} \frac{C}{i}\right) = \frac{1}{rH_N}, \quad H_N = \sum_{i=1}^{N} \frac{1}{i} \qquad [1]$$

Empirical tests (18) have shown Zipf's law to be an excellent fit for word frequencies across languages and genres, especially for relatively common words (e.g., the top 10,000 words in English). Zipf's law implies that much of the probability mass in a linguistic sample comes from relatively few but highly frequent types. Many words, over 40% in the Brown Corpus, appear only once in the sample. It is also worth noting that Zipf's law is not unique to language and has been observed in many natural and social phenomena (reviewed in ref. 19).

I now investigate how Zipf's law affects the combinatorial diversity in language use. Consider the English noun phrase (NP) rule in previous studies of child language (10–12), where a closed class functor ("a" and "the") can interchangeably combine with an open class noun to produce, for example, "a/the cookie" or "a/the desk." These combinations can be described as a rule "NP→DN," where the determiner (D) is "a" or "the" and the noun (N) is "car," "cookie," or "cat," for example. Other types of grammar rules can be analyzed in a similar fashion.

Two empirical observations are immediate (details are provided in *SI Text*). First, because many nouns appear only once, as indicated by Zipf's law, they can only be used with a single determiner. The lack of opportunities to be paired with the other determiner has been inappropriately interpreted as a restriction on combinations, as Valian et al. (12) noted. Second, even when a noun is used multiple times, it may still be paired exclusively with one determiner rather than with both, due to entirely independent factors. For instance, although the noun phrases "the bathroom" and "a bathroom" both follow the rule NP→DN, the former is much more common in language use. By contrast, "a bath" appears more often than "the bath." These use asymmetries are unlikely to be linguistic but only mirror life. Empirically, nouns favor one determiner over the other by a factor of 2.5, which is also approximated by Zipf's law. Thus, even if a noun appears multiple times in a sample, there is a still significant chance that it will be paired with one determiner exclusively. Taken together, these statistical properties of language may give rise to low-syntactic diversity measures, which have been interpreted as memorization of specific strings of words. At the same time, the Zipfian characterization of language provides a simple and accurate way to establish the statistical profile of grammar.

Statistical Profile of Grammar. Keeping to the determiner-noun example, I calculate the expected ratio of nouns combined with both determiners in a linguistic sample. The two categories of words and their combinations can be likened to the familiar urns and marbles in probability problems. Consider two urns: One contains a red marble and a blue marble, the other contains N distinct green marbles, and all the marbles are drawn with fixed probabilities. A trial consists of independently drawing one marble from each urn (with replacement); that is, a green marble is paired with either a red marble or a blue marble at every trial. After S trials, where S is sample size, one counts the percentage of green marble types that have been paired with both red and blue ones. In linguistic terms, the red and blue marbles in the first urn are the determiners "a" and "the" and the green marbles in the second urn represent nouns that may be combined with the determiners.

If the pairing between determiners and nouns follows the rule NP→DN and is independent, I can calculate the expected probability of a specific DN pairing as the product of their marginal probabilities. Let n_r be the rth most frequent noun in the sample of S pairs of determiner-noun combinations and p_r be its marginal probability. Suppose the probability of drawing the ith determiner is d_i. In the S pairs of determiner-noun combinations, the expected probability of n_r being drawn with both determiners, E_r, is as follows (derivations are provided in *SI Text*):

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

In the NP case with two determiners, I have $D = 2$. The expected diversity average of the entire sample is as follows:

$$E[D] = \frac{1}{N} \sum_{r=1}^{N} E_r \qquad [2]$$

The calculation is further simplified because frequencies of words can be accurately approximated by Zipf's law; that is, the probability of a word is inversely proportional to its rank (Eq. **1**). This enables us to calculate the expected combinatorial diversity based only on the sample size S and the number of distinct nouns type N appearing in the sample.

Results

Early Child Language Follows a Grammar. The statistical analysis in Eq. **2** gives an expected ratio of nouns appearing with both determiners if their combinations are independent. The expected value can then be compared with the empirical value to see if the observed profile of use is consistent with theoretical expectation.

I evaluate 10 language samples (details are provided in *SI Text*). Nine are drawn from the publicly available data (20) of young children learning American English at the very beginning of syntactic combinations, that is, the two-word stage. For comparison, I also evaluate the Brown Corpus (13), for which the writers' grammatical competence is not in doubt. For each sample, determiner-noun pairs are automatically extracted to obtain the empirical percentages of nouns appearing with both determiners. These values are then compared with theoretical expectations from Eq. **2**. The two sets of values are nearly identical (Fig. 1*A*). Lin's concordance correlation coefficient test (21), which is appropriate for testing identity between two sets of continuous variables, confirms the agreement ($\rho_c = 0.977$, 95% confidence interval: 0.925–0.993). In other words, very young children's language fits the statistical profile of a grammatical rule that independently combines syntactic categories.

The syntactic diversities in the linguistic samples show considerable variation, highlighted by the paradoxical finding that the Brown Corpus shows less diversity than the language of young children. The nature of variation is formally analyzed in *SI Text*, which suggests that the average number of times a noun is used in the speech sample (S/N) predicts the diversity measure (a similar analysis is presented in ref. 12). This prediction is strongly confirmed, because the average number of occurrences per noun correlates nearly perfectly with the diversity measure ($\rho = 0.986$, $P < 10^{-5}$). As the results from the Brown Corpus show, the previous literature is mistaken to interpret the value of combinatorial diversity as a direct reflection of grammatical productivity (6, 11).

Role of Memory in Language Learning. I now ask whether children's determiner use can be accounted for by models that memorize specific word combinations rather than using general rules, as previously suggested (9). To test this hypothesis, I consider a model that retrieves from jointly formed, as opposed to productively composed, determiner-noun pairs. In contrast to the grammar model, which can be viewed as drawing independently from two urns, the memory model is akin to (invisible)

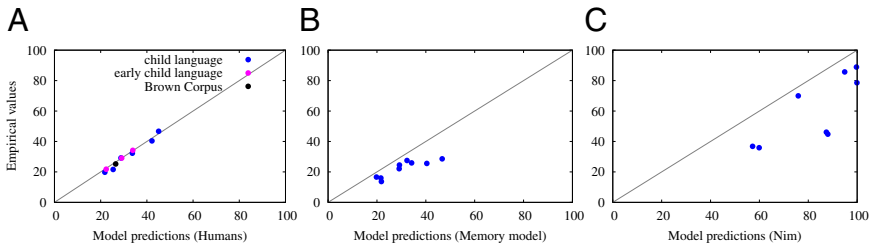

Fig. 1. Syntactic diversity in human language (*A*), memory-based learning model (*B*), and Nim Chimpsky (*C*) (details of the data are provided in *SI Text*). The diagonal line denotes identity; close clustering around it indicates strong agreement. For humans and Nim, the model predictions are made on the assumption that category combinations are independent. For the memory-based learner, the model prediction is based on frequency-dependent storage and retrieval. Only human data are consistent with a productive grammar ($\rho_c = 0.977$). Both the memory-based learning model ($P < 0.002$) and Nim ($P < 0.004$) show significantly lower diversity than expected under a grammar.

strings connecting balls from the two urns: Drawing a noun can only be paired with the determiner(s) that the learner has observed in the input data.

The memory model is constructed from 1.1 million child-directed English utterances in the public domain (20), which is approximately the amount of input children at the beginning of the two-word stage may have received (22). This sample of adult language contains nouns that appear with both determiners, as well as a large number that appear with only one determiner, for reasons discussed earlier.

For each of the nine child language samples, I randomly draw a matching number of pairs (with replacement) from the memorized determiner-noun pairs; the probability with which a pair is drawn is proportional to its frequency. I calculate the ratio of nouns in the drawing that appear with both "a" and "the" over 1,000 simulations and compare it with the empirical values of multiple determiner-noun combinations (*SI Text*). Results (Fig. 1*B*) show that the model significantly underpredicts the diversity of word combinations for every sample of child language ($P <$ 0.002, paired one-tailed Mann–Whitney test).

There is no doubt that memory plays an important role in language learning: Words and idioms are the most obvious examples. Our results show that memory cannot substitute for the combinatorial power of grammar, even at the earliest stages of child language learning. Moreover, because the memory model produces significantly lower diversity measures than the empirical values, children in the present study are unlikely to be using a probabilistic mixture of grammar and memory retrieval. That would only produce diversity measures between the memory model and the grammar model, the latter of which, alone, already matches the empirical values accurately.

Nim's Sign Combinations. What about our primate cousins? I investigate the sign combinations of Nim Chimpsky, a chimpanzee who was taught ASL (7, 15). Nim acquired ∼125 ASL signs and produced thousands of multiple sign combinations, with the vast majority being two-sign combinations. Like the NP rule NP→DN, Nim's signs can be described as a closed class functor, such as "give" and "more," combined with an open class sign, such as "apple," "Nim," or "eat." The data include eight construction types that could be viewed as potential rules (7) (complete descriptions are provided in *SI Text*). These combinations are not fully language-like, because the signs in these constructions do not fall into conventional categories, such as nouns and verbs. They do, however, suggest Nim's ability to express meanings in a combinatorial fashion.

I calculate the expected diversity from Eq. **2** if Nim followed a rule-like system that independently combines signs. Given the relatively small number of open class items and large number of combinations, a productive grammar is expected to have very high diversity measures. I then compare these expected values against the empirical values in Nim's sign combinations. Results (Fig. 1*C*) show that Nim falls considerably below the expected diversity of a rule-based system ($P < 0.004$, paired one-tailed Mann–Whitney test). Our conclusion is consistent with the video analysis results that Nim's signs followed rote imitation rather than genuine grammar (7).

Nonhuman primates' communicative capacity is complex, and their ability to learn word-like symbolic units is well documented (16, 23). There might be other cognitive systems that differentiate between humans and nonhuman primates (24) and play a role in the development and evolution of language. Our result is a rigorous demonstration that Nim's signing lacked the combinatorial range of a grammar, the hallmark of human language evident even in very young children's speech.

Discussion

I envision the present study to be one of many statistical tests to investigate the structural properties of human language. This represents a unique methodological perspective distinct from most behavioral studies of language and cognition, which typically rely on differentiation between experimental results and null hypotheses (e.g., chance level performance). By contrast, the test developed here produces quantitative theoretical predictions, where one seeks statistical confirmations, rather than mismatches, against empirical data. Zipf's law, which provides a simple and accurate statistical characterization of language, enhances the robustness and applicability of the test across speakers and genres.

The present study also helps to clarify the nature of children's early language. It suggests that at least some components of child language follow abstract rules from the outset of syntactic acquisition. I acknowledge the role of memorization in language (6), but our results suggest that it does not fully explain the distributional patterns of child language. This conclusion is congruent with research in statistical grammar induction and parsing (25). Statistical parsers make use of a wide range of grammatical rules. The verb phrase (VP) "drink water" may be represented in multiple forms, ranging from categorical (VP→V NP) to lexically specific (VP→V_{drink} NP) or bilexically specific (VP→V_{drink} NP_{water}), which corresponds to specific word combinations suggested for child language. When tested on novel data, it has been shown that generalization power primarily comes from categorical rules and that lexically specific rules offer very little additional coverage (26). Taken together, these results suggest that in language acquisition, children must focus on the development of general rules rather than the memorization and retrieval of specific strings.

Finally, the quantitative demonstration that children but not primates use a rule-based grammar has implications for research into the origin of language. Young children spontaneously acquire rules within a short period, whereas chimpanzees appear to show only patterns of imitation even after years of extensive training. The continuity between the ontogeny and phylogeny of

language, frequently alluded to in the gradualist account of language evolution (2–5), is not supported by the empirical data. Theories that postulate a sharp discontinuity in syntactic abilities across species (27) appear more plausible.

1. Chomsky N (1957) *Syntactic Structures* (Mouton, The Hague).
2. Bickerton D (1996) *Language and Human Behavior* (Univ of Washington Press, Seattle).
3. Wray A (1998) Protolanguage as a holistic system for social interaction. *Language and Communication* 18(1):47–67.
4. Studdert-Kennedy M (1998) *Approaches to the Evolution of Language: Social and Cognitive Bases*, eds Hurford J, Studdert-Kennedy M, Knight C (Cambridge Univ Press, Cambridge, UK), pp 202–221.
5. Hurford J (2011) *The Origins of Grammar: Language in the Light of Evolution* (Oxford Univ Press, New York).
6. Tomasello M (2003) *Constructing a Language* (Harvard Univ Press, Cambridge, MA).
7. Terrace HS, Petitto LA, Sanders RJ, Bever TG (1979) Can an ape create a sentence? *Science* 206(4421):891–902.
8. Brown R (1973) *A First Language* (Harvard Univ Press, Cambridge, MA).
9. Tomasello M (2000) First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11(1–2):61–82.
10. Valian V (1986) Syntactic categories in the speech of young children. *Dev Psychol* 22(4):562–579.
11. Pine J, Lieven E (1997) Slot and frame patterns in the development of the determiner category. *Appl Psycholinguist* 18(2):123–138.
12. Valian V, Solt S, Stewart J (2008) Abstract categories or limited-scope formulae? The case of children's determiners. *J Child Lang* 35(1):1–36.
13. Kučera H, Francis N (1967) *Computational Analysis of Present-Day English* (Brown Univ Press, Providence, RI).
14. Anderson SA (2006) *Doctor Dolittle's Delusion: Animals and the Uniqueness of Human Language* (Yale Univ Press, New Haven, CT).
15. Terrace HS (1979) *Nim* (Knopf, New York).
16. Savage-Rumbaugh ES, Shanker S, Taylor TJ (1998) *Apes, Language, and the Human Mind* (Oxford Univ Press, New York).
17. Zipf GK (1949) *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison–Wesley, Reading, MA).
18. Baroni M (2009) *Corpus Linguistics*, eds Lüdelign A, Kytö M (Mouton de Gruyter, Berlin), pp 803–821.
19. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46(5):323–351.
20. MacWhinney B (2000) *The CHILDES Project* (Lawrence Erlbaum, Mahwah, NJ).
21. Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45(1):255–268.
22. Hart B, Risley TR (1995) *Meaningful Differences in the Everyday Experience of Young American Children* (Brookes, Baltimore).
23. Lyn H, Greenfield PM, Savage-Rumbaugh S, Gillespie-Lynch K, Hopkins WD (2011) Nonhuman primates do declare! A comparison of declarative symbol and gesture use in two children, two bonobos, and a chimpanzee. *Language and Communication* 31(1):63–74.
24. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind. *Behav Brain Sci* 1(4):515–526.
25. Collins M (2003) Head-driven statistical models for natural language processing. *Computational Linguistics* 29(4):589–637.
26. Bikel D (2004) Intricacies of Collins parsing model. *Computational Linguistics* 30(4): 479–511.
27. Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598):1569–1579.