Who's afraid of George Kingsley Zipf?

Or: Do children and chimps have language?

Every parent hears their child's first words with pride. Phrases like "A doggie" soon follow. Or is it sometimes "The doggie"? The difference sheds light on when children (or chimps) first acquire true language – once we have taken Zipf's law and yodelling turtles into account. **Charles Yang** explains.

The Belgian writer Georges Simenon, with some 200 books to his credit, attributed his prodigious output to the verbal deficiency of his readers. Most French people, Simenon claimed, used only 600 words, so his writing had to be simple and straightforward. It is easy to dismiss Simenon's numerical sense: this is the same author who claimed to have slept with 10 000 women. Working out how young children acquire language, and whether chimpanzees can achieve the same trick, has a surprising connection with Simenon's 600-word claim.

In fact, Simenon may not have been that far off – so far as words are concerned. Supporting evidence comes from across the Atlantic. When studying the vocabulary use of James Joyce, the Harvard linguist George Kingsley Zipf made a startling discovery in the 1930s about the statistics of language. When we speak or write, we use a relatively small number of words very frequently, while the majority of words fall on a very long tail and are hardly used at all.

In the current Big Data age, it is easy to verify Zipf's findings. Plotting the distribution of the top 1000 words in a million-word collection of English writings, we immediately see the dramatic decline of frequency, followed by a long and essentially flat tail where the majority of words reside. Figure 1 shows what happens. "The", "of", and "and" appear tens of thousands of times; most other words are lucky if they make it once or twice.

More precisely, and in what has become known as Zipf's law, the rank of words and their frequencies are inversely related, with their product approximately a constant. The most frequent word in our collection, of course, is "the", which occurs almost 70 000 times. The second most frequent word is "of", bearing the rank of 2 and appearing just over 36 000 times. Now 72 000 (2 × 36 000) is not exactly 70 000 but is awfully close. The inverse relationship between rank (r) and frequency (f) is best viewed on a log scale. If $rf = C$, where C is some constant, then $\log f = \log C - \log r$: the rank and frequency would form a straight line with a slope of −1. The top 1000 words are shown again in Figure 2, this time plotted logarithmically. Linear regression fits a slope of −0.98, nearly perfectly confirming Zipf's Law.

The most frequent word in the English language is "the". It occurs twice as often as the runner-up, which is "of"

It turns out that the probability mass of language is largely concentrated in the upper echelon of words. There are about 50 000 unique words in our million-word-altogether collection, but the top 130 of them – or just 0.3% – appear so often that they already account for 50% of all uses, while 45% of all words appear exactly once. An economic example helps illustrate just how incredibly unfair the lives of words are. In 2012, the top 1% claimed almost 20% of all household income in the US; it provoked public outrage and became the allegory of the polarisation of wealth. Words are far more top heavy: the top five alone ("the", "of", "and", "to", "a") account for 20% of words in English writings. Most of the words, just like most of us, are among the 99%. Simenon only slightly overstated his case: the top 600 words make up 70% of French writings.

It remains unclear why Zipf's law should hold in language after language and study after study. Don't rush your answer: intellectual heavyweights have battled out over the years without coming to an agreement. On the one hand, Zipf's law, and its cousins in the family of power law distributions (where one quantity varies as a power function of the other, Zipf's law being the simplest), pop up all over the place. According to the 2010 census, New York, the most populous city in the US, had 8.2 million residents, roughly doubling the second ranked Los Angeles (3.8 million), tripling the third ranked Chicago (2.7 million), quadrupling the fourth ranked Houston (2.1 million) and quintupling the fifth ranked Philadelphia (1.5 million). Coincidence? Maybe not, but no one knows for sure. Paul Krugman, the Nobel laureate in economics and *New York Times* columnist, once lamented: "The usual complaint about economic theory is that our models are over-simplified – that they offer excessively neat views of complex, messy reality. … in one important case [Zipf's law] the reverse is true: we have complex, messy models, yet reality is startlingly neat and simple." At the same time, and at least equally surprisingly, random combinations of letters also follow Zipf's law, often providing an even better fit than real words. George Miller, the psychologist most famous for demonstrating that the short-term human memory has an average limit of 7 digits, proved that a monkey randomly banging on a keyboard, intermittently striking the spacebar, will generate "words" that follow Zipf's law almost to the letter. (He did it by a statistical argument, not with real monkeys.)

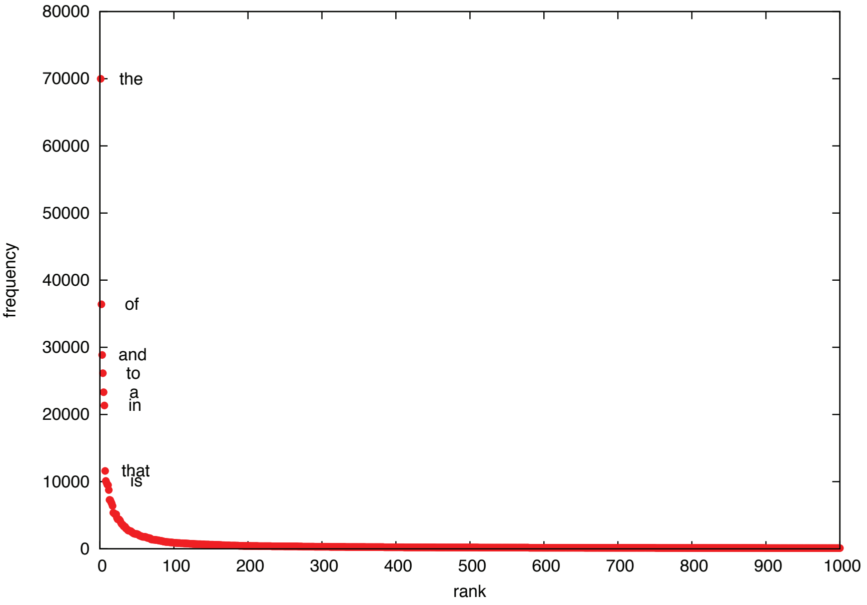

Figure 1. In a million words of writing in English, the word "the" appears 70 000 times, "of" appears about half as often, and most words occur just a few times or only once

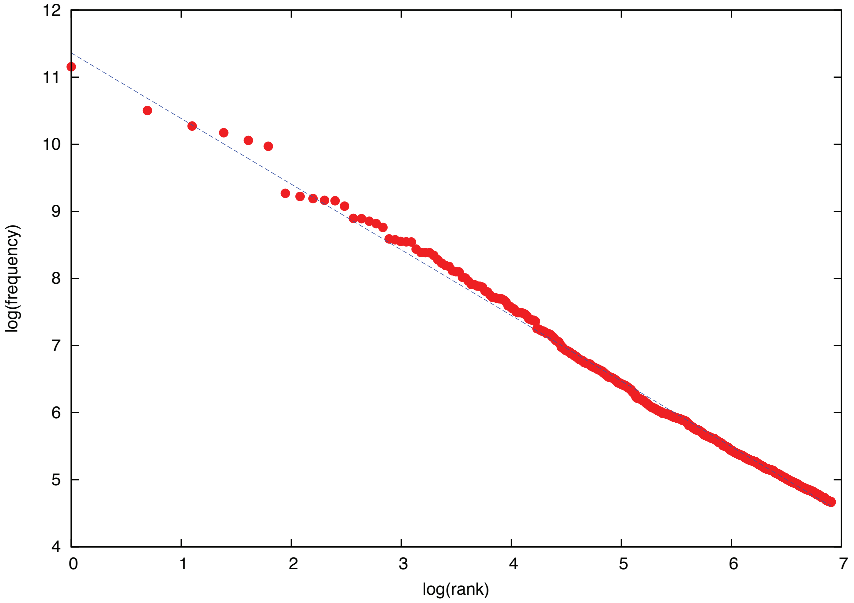

Figure 2. The word frequencies of Figure 1 plotted logarithmically

In all likelihood, Zipf's law will not hold the secret of language, never mind cities and the market force. My own theory is that humans are boring, and we keep talking about the same thing. (Evidently we keep asking the same thing too: the frequencies of internet queries also follow Zipf's law.) Our undertaking here is much more modest: Zipf's omnipresence frustrates as well as enlightens the study of language.

Zipf checkmates Google?

Zipf is Google's worst nightmare.

"Nine and a half turtles wearing orange aprons with watermelons on them began to simultaneously yodel as the conductor stepped on the kitchen table." I am certain you have never heard this sentence before, yet every speaker of English instantly understands its meaning, however unsettling the imagery may be. The hallmark of language is its infinite range of forms and meanings, ultimately rooted in the boundless creativity of human thought and imagination. As the world's leading purveyor of information, it is Google's job to find keywords, extract meanings, and translate Arabic into Zulu. The foremost weapon in Google's arsenal is the astronomically large collection of language data that is expanding at an ever faster pace, but even Google has not come across our nine and a half yodelling turtles.

Most modern natural language processing systems analyse sentences by breaking them into a sequence of chunks of two, three, four words and so on. The probabilities of these chunks are individually estimated from how often they occur in that massive database before they are strung back together, usually by means of multiplication, to determine the probability of the entire sentence. One serious challenge here is that, as the size of the chunks grow, so does the space of word combinations. For a vocabulary of N words (say 10^5, as is in English), there are N^2 pairs, N^3 triples, N^4 quadruples, with an exponential growth that may prove too fast even for Google. The problem is ameliorated partly by the fact that not all combinations of words are possible: "nine and a half turtles" is grammatical, however unlikely, but scrambling them around into "turtles half nine a the" makes a word salad that neither an English speaker nor Google cares for.

When my four-year-old daughter heard about the nine and a half yodelling turtles, she gave me a funny look: "That's silly!"

This is where Zipf's law rears its ugly head. In an egalitarian non-Zipfian world, where words and their combinations are uniformly distributed, we could expect to witness all of the admissible word combinations, at least for shorter chunks, given the volume of data that pours into Google's data collection every second. After a while, we may safely conclude that the missing combinations will never appear because they are impossible: absence of evidence can substitute for evidence of absence. In the actual world, however, we are undercut by Zipf's long tail, which grows longer as the space of word combination gets larger. In our million-word collection of English, 45% of words, 80% of word pairs, and a full 95% of word triples appear exactly once. Past research has shown that as new data comes in, so do new chunks of words. "Turtles wearing orange aprons" was not in Google's database last week; but it is there now, thanks to this article.

Nine and a half turtles wearing orange
aprons with watermelon them began
to simultaneously yodel as the conductor
stepped on the kitchen table

Illustration: Andrew Tapsell (www.andrewtapsell.com)

Here is the dilemma for Google: how should it treat combinations – such as "nine and a half turtles" and "turtles half nine a the" – that have not (until just now) made an appearance? The very imbalanced Zipfian use of language implies that many grammatical strings will not appear anytime soon, which makes them difficult to distinguish from ungrammatical ones that will never appear.

Considerable ingenuity and engineering know-how have gone into overcoming this problem. Now Google Translate can render restaurant reviews from Finnish into English: well, almost English, but at least the traveller will not go hungry. Yet natural language processing still has a long way to go, especially as it tries to move beyond strings of words into the realms of meanings and intentions.

For instance, asking Google "Which Starks died in Game of Thrones" gives you many informative hits but negating the query ("Which Starks didn't die …") gives you pretty much the same thing.

A long time has passed since Deep Blue defeated Gary Kasparov, and futurists have been telling us that singularity – that point at which machines become cleverer than people and the future changes beyond guessing – is just around the corner. It is perversely comforting to know that all the computing power in the world still stumbles over the complexity of human language. This is more impressive than it seems: not everyone learns chess and almost no one will ever approach the level of Kasparov, but everyone learns a language effortlessly. And we do so on much less data than Google. Children acquire the essential components of language in a few short years, with no more than several million sentences many of which are, as you would expect, repetitive uses of "Careful!", "What's that?" and "I don't know". Zipf's law, a universal property of language, needs to be overcome by young children and Google alike. When my four-year-old daughter heard about the nine and a half yodelling turtles, she gave me a funny look:

"That's silly!"

"A", "the", and infinite variation

Birds sing, kids chatter; language is our destiny, a gift from evolution. But babies are not born talking: at some point of their development, language just clicks in and switches on. Like curious parents, scientists also want to know when language starts.

The answer partly depends on what we take language to be. There is now evidence that even newborns react to the rhythm of their native language. Three-year-olds can carry on charming conversations, while acing the SAT Vocabulary Test for university entrance requires a very disciplined teenager. But there is broad agreement that the heart of language is the ability to combine words to express infinitely varied meanings, and that seems to be absolutely unique among the communication systems in the biological world: I am communicating to you how we communicate, something no other animal does.

It is not easy to figure out when children really start putting words together, especially if we wish to push the envelope, so to speak, to the very earliest stage of language learning. There is an oft-repeated story about Einstein as a very late talker. "The soup is too hot", as legend has it, were his first words, at the very ripe age of 3: apparently the boy genius had not seen anything worth commenting on before that. The credulity of such tales aside – similar stories abound with other famous subjects – they do contain a kernel of truth: a child does not have to say something, *anything*, just because he can. This poses a major obstacle for the study of child language learning. We may turn to clever experiments, capitalising on the child's interest in blinking lights, moving objects and talking puppets. But these are no substitute for unhindered natural speech when children interact with their mothers, which is often the only, and certainly the most accessible, data on hand.

When young children do talk, they are often quite repetitive (call them strong-willed if you prefer): "a doggie", "a doggie", "a doggie". They may get their wish in the end, but is this truly language? Recall the essence of language is the infinite diversity of linguistic combinations. Take the simplest combination of words in English: a noun phrase that is composed of an article ("a" or "the") and a singular noun (such as "doggie"). Since the two articles are interchangeable, one might expect a noun paired with one to be automatically extended to the other, which would be the benchmark of linguistic proficiency. But young children fall short: in their speech samples, only 20–40% of nouns that appear with one also appear with the other. If you listen to a two-year-old for a few hours, repetitive use of "a doggie" may be the only context in which "doggie" makes an appearance. "The doggie" never seems to happen.

The ability to learn grammar may require several years to emerge, as some researchers conclude: the only way children know how to talk about doggies is to say "a doggie", perhaps as the result of rote memorisation of their parents' language. So a child who says "a doggie" without also saying "the doggie" has not yet acquired language. But then we would be facing some paradoxical results. The psycholinguist Virginia Valian and her students find that applying the same metric of linguistic diversity to the speech of *mothers* yields low measures comparable to those of their children, yet the mothers' linguistic sophistication is not under question. More puzzling still, in our million-word collection of English writings, only 25% of singular nouns that appeared with one article also appear with the other, leading to a diversity measure lower than that of some young children – who, as we have seen, can put "a" and "the" with up to 40% of their nouns. Some two-year-olds, then, would have a better command of the English grammar than professional writers — which seems absurd.

Suppose you never need to say "The doggie"?

In most social and behavioural sciences, statistics are used to disprove the null hypothesis: to see whether the test score differences between two demographic groups are real, or whether using cell phones has any ill effects on health. These tools also feature in the study of language and child development. For instance, we may conclude that the baby has grasped some aspects of a word's meaning if she looks at the matching picture significantly longer than at some non-matching distractor. The key to all good statistical tests is to develop a well-informed null hypothesis that is worth rejecting.

Suppose you were dealt ten hands of poker and not even a pair showed up. Before claiming that the dealer is bent we need to reject the null hypothesis of an unlucky streak. Language is also a game of chance. Words such as articles and nouns are like cards in a deck; we draw a fresh hand every time we speak. The child might not have drawn enough pairs, but more could be just be around the corner with a few more hands. To complicate the matter further, the cards are in a game of equal opportunities, having a uniform probability of being drawn. Words are not. To work out the odds for the game of language, Zipf's law needs to be part of the null hypothesis.

Let us say the children have perfect command of the grammar, whereby they combine the articles and nouns completely interchangeably and independently. They say "a doggie" and "the doggie" as occasion requires. Two factors conspire to account for the surprisingly low level of grammatical diversity. The first is obvious. Many of the nouns paired with articles will have occurred only once, somewhere on Zipf's long tail; our child might say "the giraffe" on a trip to the zoo but have no other opportunity to mention giraffes, either as "a" or "the"; if you appear only once, you cannot be paired with both articles, and you will bring down the average.

The other reason is less obvious, and in turn demonstrates the depth of Zipfian reach. When you toss a coin five times, you might get a mix of heads and tails, or five tails in a row. The latter may not be very likely – if the coin is fair. For a noun, its pairing with "a" and "the" is similar to a coin toss, except that the coin is very heavily biased. Nouns tend to have a favoured article even though both combinations are possible. For instance, "the bathroom" is more commonly used than "a bathroom", but we say "a bath" a lot more often than "the bath". All four phrases follow the rule of English but the imbalance in their usage surely is not a matter of grammar, which presumably does not control personal hygiene. On average, the more favoured article ("the bathroom") more than doubles the less favoured one ("a bathroom"), much as the most frequent word dominates the next most frequent one – *à la* Zipf.

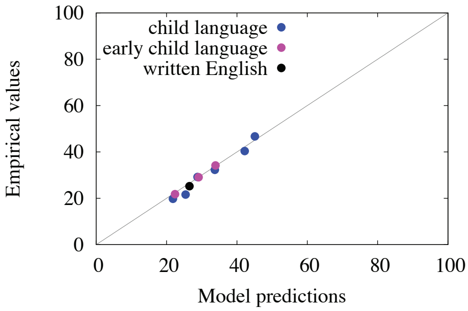

Figure 3. The language of young children fits the diversity predicted by a model based on the statistics of adult language.

What remains is an exercise in probability theory, similar to working out the odds for actual poker hands. We want to test the hypothesis that the child uses "a" and "the" combinations in the proportions of language ("grown-up" language if you like) as opposed to rote. At each trial, a noun ("bathroom", "doggie") is selected according to its frequency. The metaphorical coin is tossed, landing on one of the two articles, "a" or "the", not with equal probability but with a Zipfian favourite. For each noun that appears in a collection of such phrases, we can calculate its expected probability of having been paired with both articles. Here Zipf's law makes life easy. We do not need to empirically estimate the frequencies of each noun from a speech sample: if there are N unique words, Zipf's law tells us that the rth ranked noun will have the following probability of being used:

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

since its rank is inversely proportional to its frequency. We do not even have to know what these nouns are. Surely different children use different nouns: for one child, the most frequent noun may be "truck", and for another, "doggie". But as long as noun frequencies follow Zipf's law – and they do – we can calculate the expected diversity of article–noun pairings using only N and the sample size. We can compare it to the diversity that the children actually show.

The toddler deserves more credit than we do. I took nine speech transcripts of children learning to speak English. The data was gathered at the very beginning stage of putting multiple words together, with three youngest samples starting at just under 2 years of age. To calibrate the model, I also tested the expected diversity for the million-word collection of written English: if the calculation is correct, it should accurately predict the behaviour of professional writers whose linguistic competence is not in doubt.

The result? Agreement between the theoretical and empirical values is very tight. "A" and "the" in child language, even in early child language, closely correspond to their use in written English. Indeed, the usual statistical tests for significant differences fail to distinguish them. It seems our young children have already acquired real language.

Zipf's law, which enables a precise characterisation of language, has given us a rare opportunity in the study of human behaviour to look for agreement between theory and reality. How firm is our conclusion? The more typical statistics in social sciences is predicated on rejecting the null hypothesis (e.g., "chance-level performance"). This is useful but not as convincing as empirically confirming a mathematically derived prediction: the physicist does not declare victory when the results coming out of the particle accelerator are statistically significantly different from random noise.

Here we turn to the toolkit from biomedical statistics. A drug is reliable if its efficacy is sufficiently similar across a range of cases. To test for agreement between theoretical and empirical results, we view them as two separate clinical trials, with each case representing a sample of language data. The concordance correlation coefficient, which was designed for measuring reproducibility, confirms the agreement. If predictions and experiments agree – to the eighth digit after the decimal point, as they do in physics – the coefficient will be 1.0, and the two sets of values will fall on the identity line. As it happens, we are not very far off: the coefficient is ρ_c = 0.977 (95% confidence interval 0.925–0.993), and the points cluster around identity closely (Figure 3). In other words, the diversity of language usage by

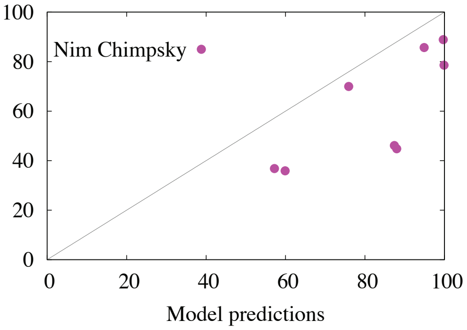

Figure 4. Nim Chimpsky did not achieve the diversity of language; he did not have a true grammar

A doggie, the doggie or my doggie? Ingram Publishing/Thinkstock children (and, reassuringly, adults) is exactly what one might expect from a grammar rule, once the general statistical properties of language are taken into account. Young children have real proper language, and we can say so with some certainty.

Maybe this is all obvious to the perceptive parent. Still, it is good to have some solid statistical proof. To be sure, the two-year-old still has a long way to go: words, idioms, complex phrases, the appropriate way to talk to friends, teachers, and in-laws (which takes a lifetime). But children clearly have a biological predisposition for language learning: a little bit of data, even when heavily skewed by Zipf's law, suffices to set the language engine in motion, to the envy and marvel of technologists.

Chimps chatter?

We wrap up with a parallel case that we can use as a comparison: did Nim Chimpsky, the signing chimpanzee, have language? The fascination with talking animals goes as far back as King Solomon, who was said to be able to converse with animals, and it has continued to occupy a central place in the study of language and its biological place in nature. Non-human primates seem to have some precursory abilities for language, or at least for some of the major components of language. For instance, the ability to acquire arbitrary associations between forms and meanings has been amply demonstrated in primates and other species, although important differences with child word learning may still remain. Dolphins have been trained to associate symbols with objects, for example; but this is not language. The defining property of language is the free and unbounded combination of linguistic units to express meanings; and evidence for this in animals is far more equivocal.

Signing apes have been at the forefront of animal language research, captivating the general public as well as the scientific community. Nim Chimpsky was a chimpanzee reared in a human family from infancy, with systematic instructions in American Sign Language. (He also rode in cars, learned to use the toilet, and smoked marijuana: this was the 1970s.)

To this day, Nim has provided the only public database of signs from animal language studies. (By contrast, the ability of Koko, the famous talking gorilla who occasionally holds online chats, comes exclusively from her trainer's interpretation and YouTube clips.) Nim learned some 125 signs of American Sign Language, which was pretty much par for the course for chimpanzees. Far more impressive is his multiple sign productions, some 20 000 in all: "give Nim", "give banana", "more banana", "more water". Herbert Terrace, the lead investigator of the project, initially thought that Nim had crossed the great linguistic divide in the primate order. He appeared to use "give" and "more" as a template, where the open slot which follows can be filled in by a wide range of signs. Terrace was able to identify eight robustly attested templates. If genuine, these templates are fundamentally the same system as the rule for English noun phrases, with the articles "a" and "the" acting as holding places for a wide range of nouns. Further examination of Nim's signing videos, however, convinced Terrace otherwise. Nim tended to model immediately after his teacher's signs, with very few instances of spontaneous combination. A high-profile report appeared in *Science*, and its negative conclusion essentially ended animal language projects (and more importantly, funding opportunities).

With our statistical profiler for grammar, we can turn to the 40-year-old data on Nim for a post mortem. Nim showed Zipfian tendencies in his communications: not all signs were used equally frequently, with "me", "Nim", "drink", "banana", "apple" among the predictable favourites, while most signs are used rarely, many only once. What if Nim had a rule-like system, one which swapped "more" and "give" interchangeably to combine with "banana", "ball", and "water"? The degree of diversity, or the percentage of signs that appear with both "more" and "give", ought to match the theoretical calculation used to verify the presence of grammar rules in human speech. We put Nim's data through the same kind of analysis that we used for two-year-old children; and it looks as though Terrace was right. Nim's signs fall significantly below the expected benchmark for a grammar (Figure 4); for this, the conventional statistical tests for differences are just the right tools ($p < 0.004$). Nim's combinations, of "more" and "give" with differing nouns, show significantly lower diversity than in a language with grammar. He was memorising, not learning a rule.

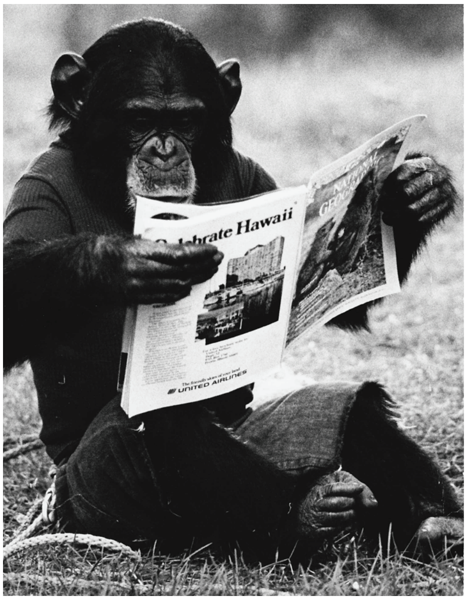

Photo: Herbert Terrace

Nim's life took a turn for the worse after the sign language study. He was unable to relate to other chimps, was transferred to a laboratory where he was kept in a small cage, and died prematurely. (See the 2011 documentary *Project Nim*, directed by James Marsh, available on DVD.) It raises serious ethical issues when we take a highly intelligent and social animal out of its element to immerse in an utterly alien lifestyle and communication system. In all likelihood, there will never be another Project Nim; along with it goes a glimmer of hope to understand the nature and origin of language. Many scholars had high expectations of Nim. Since young children were believed to go from a stage of rote memorisation to a full-blown grammar, placing the chimpanzee's signs somewhere on this trajectory would be an ontogenetic recapitulation of phylogeny – the notion, once in vogue among theorists of evolution, that young animals go through stages resembling successive stages in the evolution of their remote ancestors. This might have shed light on how language emerged in humans' great leap forward.

Under the Zipfian light, however, the apparent continuity between chimps and children proves to be an illusion. Children have language; chimps do not. Young children spontaneously acquire rules within a short period of time; chimpanzees only show patterns of imitation after years of extensive training. The linguistic gap seems deep; otherwise chimps would have been reading this magazine.

Charles Yang is on the faculty of linguistics and computer science, and directs the cognitive science program at the University of Pennsylvania.