A. MARTIN

THE LANGUAGE BARRIER

Some researchers think that the evolution of languages can be understood by treating them like genomes — but many linguists don't want to hear about it. **Emma Marris** reports.

Consider a word as it tumbles through history: *khun*, a Nepali word for blood. In the early twentieth century, the word fell all too often from the lips of the Gurkhas, a Nepalese brigade in the British Army, in songs describing the horror of the First World War. To linguist Ralph Lilley Turner, who fought beside the Gurkhas by the Suez Canal, *khun* was one of many words transcribed phonetically for the English–Nepali dictionary he compiled in the midst of the fighting.

And then, in the 1960s, the word *khun* stopped being a word so much as a node of data in the work of linguist Isidore Dyen. Working at Yale University in New Haven, Connecticut, Dyen used Turner's dictionary to assemble a list of 200 basic Nepali word meanings, including blood, and coded them onto IBM punch cards that were fed into early computers. He used this, and similar lists from another 83 languages, to try to measure the rate at which languages change over time. His discipline of lexicostatistics is now discredited for the crude assumptions it made. But as Dyen worked on his monograph, he transferred his data on to computer disk and, in the late 1990s, published his data online[1].

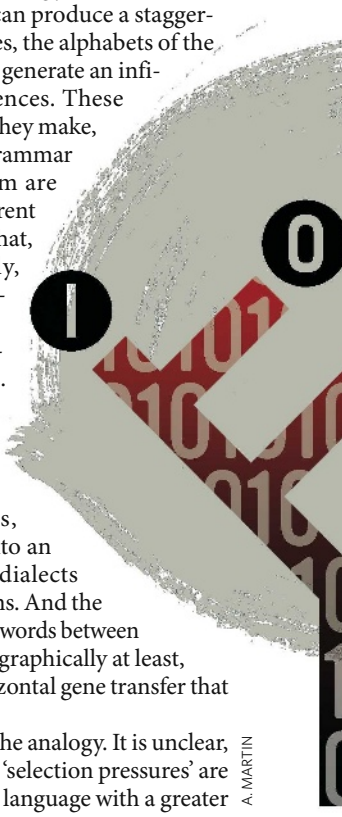
> **"We try to find mathematical patterns in nature."**
> — Martin Nowak

Nearly a century after it was sung by the Gurkhas, *khun* became a few figures of code in the computer models of evolutionary biologist Mark Pagel at the University of Reading, UK. Last year, Pagel used Dyen's online data to generate trees that estimate when languages such as Nepali diverged from one another[2]. In these models, the word is stripped of much of its rich human history. But it is exactly this type of pared-down language that speaks to researchers such as Pagel the most. "Some linguists have spent a career studying a language that becomes a single data point in one of these analyses," he says. "We do things because they are mathematically elegant, and are delighted when they can be simplified."

A new approach

In the past five to ten years, more and more non-linguists such as Pagel have used the computational tools with which they model evolution to take a crack at languages. And one can see why. Like biological species, languages slowly change and sometimes split over time. Darwin's Galapagos finches evolved either large beaks or small; Latin *amor* became French *amour* and Italian *amore*. Darwin himself noted the 'curious parallel' between the evolution of languages and species in *The Descent of Man, and Selection in Relation to Sex*.

The advent of molecular genetics provided a new depth to the analogy. Just as the four nucleotides of DNA can produce a staggering variety of creatures, the alphabets of the world's languages can generate an infinite number of sentences. These alphabets, the words they make, and the sounds and grammar rules that frame them are passed down from parent to child in a process that, at least superficially, resembles the inheritance of DNA.



A. MARTIN

Even some complications are the same. Just as species can shade off into a maddening continuum of subspecies, populations and hybrids, languages dissolve into an untidy collection of dialects and intermediate forms. And the rampant borrowing of words between languages resembles, graphically at least, the promiscuous horizontal gene transfer that microbes engage in.

There are limits to the analogy. It is unclear, for example, what the 'selection pressures' are for language, if any. A language with a greater

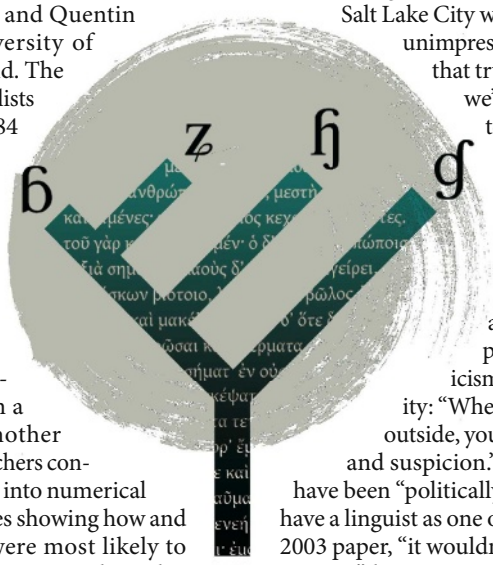
number of speakers is not obviously 'more fit' than a dying language. Although a speaker of the prevalent tongue could communicate with more people, it is not the intrinstic properties of a language that make it more widely spoken. Instead, languages seem to rise to prominence on the coat-tails of the culture that speaks them, just as the prevalence of English traces back to the broad reach of British colonialism. It's no wonder, then, that mathematical biologists such as Pagel have become interested in a system that is intriguingly like, and intriguingly unlike, genes. "I think sophisticated mathematics will increasingly become part and parcel of what we mean when we say that we have 'explained' the phenomena of language change over time," says Erez Lieberman, who studies mathematics and biology at Harvard University in Cambridge, Massachusetts.

The old school

However, there is already an old and venerable field of language-tree makers. Historical linguists have been reconstructing languages since the 1780s. Their tool is called the comparative method and it relies on extensive knowledge of the language group at hand, along with a broad grasp of, and intuitive feel for, the ways in which languages change. A linguist might notice that the way a vowel is spoken has shifted in two languages when compared with an ancient one, and infer that the shift happened before the two languages split. This will help to place the split relative to other splits but gives no information about when it happened. Hence the comparative method produces trees, but no dates.

It is putting it mildly to say that many historical linguists find the evolutionary biologists working on language histories to be bungling interlopers who have no idea how to handle linguistic data. It is also an understatement to say that some of these interlopers feel that their critics are hidebound traditionalists working on a hopelessly unverifiable system of hunches, received wisdom and personal taste. And that's just the mood between the historical linguists and the newcomers. Lots of the newcomers don't like each other either. "Why get excited about it when it is still so preliminary?" says Johanna Nichols, a historical linguist at the University of California, Berkeley. "We are not impressed by a computational or mathematical paper per se. We have to see that it blends well with what is known by historical linguistics and really adds to our knowledge. Then we will be excited."

Perhaps the most famous and controversial study[3] produced by the new school is a 2003 paper by Russell Gray and Quentin Atkinson at the University of Auckland, New Zealand. The pair started with Dyen's lists of word meanings for 84 languages from the Indian and European subcontinents, plus a few extras from extinct tongues. The data already included Dyen's opinion on which of these words were 'cognates', descended from a common word in a mother language, but the researchers converted this information into numerical code and generated trees showing how and when the languages were most likely to have branched off from one another. This same type of likelihood algorithm is used to compare species' DNA sequences and produce evolutionary trees. Specifically, Gray and Atkinson dated the origin of a language family called Indo-European to around 7,800–9,800 years ago. This ineffable date has been one of the most intensely studied and disputed points in all of historical linguistics and, based on archaeological and linguistic data, had previously been put at anywhere between about 6,000 and 10,000 years ago. When Gray and Atkinson's paper made the rounds of linguistics departments, howls of protest ensued.



A. MARTIN

"We do things because they are mathematically elegant, and are delighted when they can be simplified."
— Mark Pagel

Some critics took the paper as a return to glottochronology, a discredited method from the middle of the twentieth century — and cousin of Dyen's lexicostatistics — which in most cases disastrously assumed that all languages change at a constant rate and which helped turn linguists against any quantitative analysis of their treasured subject. But Gray and Atkinson's statistical method does not assume uniform rates of change. Many historical linguists also felt that similarities between words are a terrible proxy for similarities between languages. They tend to argue that common sounds and grammatical rules are stronger evidence for common descent than individual words, which may be similar due to chance, borrowing, or even 'nursery formations' such as mama and dada — words that mirror each other simply because all infants babble similar things.

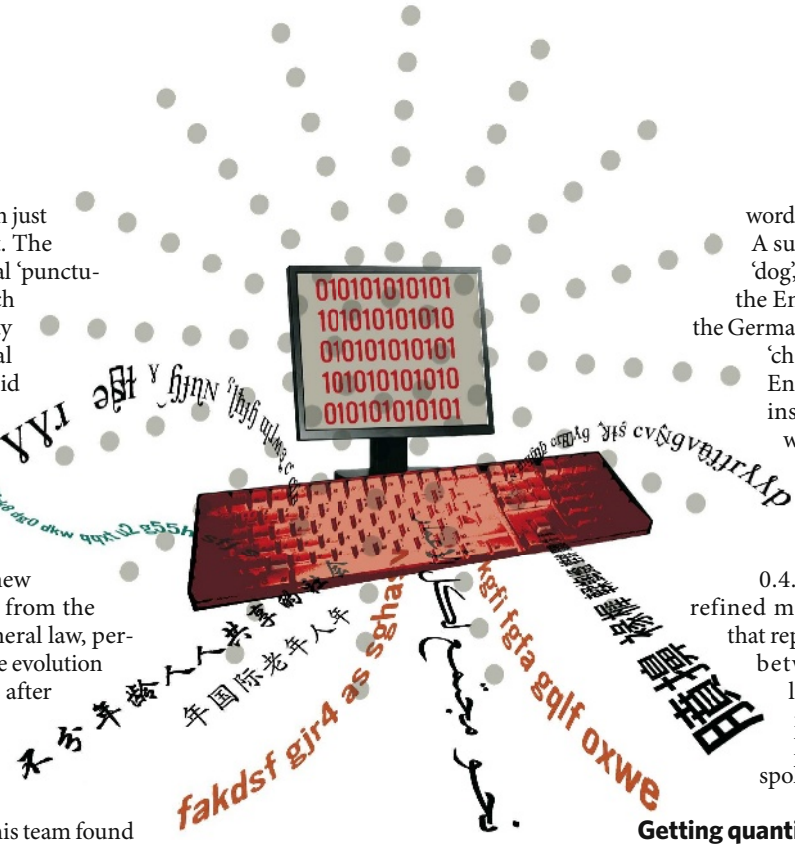
"I think that some of these researchers think that these analyses are going to supplement or even supplant historical linguists," says Lyle Campbell, a linguist at the University of Utah in Salt Lake City who was one of those unimpressed. "So far, the ones that try to go beyond what we've done don't seem to work."

Gray says that the tree does work even though it doesn't take into account the subtleties of sounds and grammar, and he puts much of the criticism down to territoriality: "When people come from outside, you see a bit of hostility and suspicion." Although it might have been "politically more palatable" to have a linguist as one of the authors on the 2003 paper, "it wouldn't have changed the answers", he says.

Ultimately, many linguists felt that this type of analysis oversimplified their cherished subject more than they could bear. Linguists love the little details that give a language personality: to them, the identifying sounds or peculiar borrowed words are nuances that tell the tale of a tongue. The new breed brushes over these details in pursuit of generalities, trends and statistical rules. "We try to find mathematical patterns in nature," says Martin Nowak, an evolutionary modeller at Harvard. "If someone works on the detailed classification, they might be dissatisfied with something that is cruder."

Grand ambitions

That dissatisfaction looks set to grow as many in the new school pursue grander ambitions: to find quantitative laws that describe language evolution. In a recent example, published in *Science* earlier this year, Pagel, Atkinson and their colleagues used word lists to build trees in three of the world's major language families; Indo-European, Bantu (an African language family that includes Swahili) and Austronesian — spoken on Pacific Islands[4]. They found that between 10% and 33% of divergence among these languages happened in what they called 'punctuational bursts', phases



A. MARTIN

of accelerated language evolution just after a language splitting event. The finding echoed the controversial 'punctuated equilibrium' theory in which Niles Eldredge and Stephen Jay Gould proposed that biological evolution often occurred in rapid bursts amid longer periods of relatively slow change. Pagel and his team speculate that the bursts could arise from the spoken idiosyncrasies of a small number of population founders, or a desire within a new population to sound different from the other group. So here is one general law, perhaps: up to one-third of language evolution occurs in punctuational bursts after splitting events.

A second possible law arose from studies of word frequency. In their 2007 study, which was published in *Nature*, Pagel and his team found that 50% of the difference in language evolution rates could be explained by the frequency with which words within the language were being used[2]. Often-used words were 'stickier' and resisted change. "What really excites me about the frequency effect is that we are identifying a general evolutionary law," Pagel says. "We think it will hold and will have held since we began talking."

In the same issue of *Nature*, Lieberman, Nowak and their co-workers showed that irregular English verbs become regularized more quickly if they are rarely used[5]. So the past tense of a rare verb such as 'gnaw' would have a 50% chance of regularizing to 'gnawed' from the Old English form 'gnagan' in 700 years. By contrast, a very common verb such as 'be' would have a 50% chance of regularizing to 'beed' in 38,800 years, perhaps explaining why 'was' remains the preferred form today. The researchers even had a precise mathematical description of the trend: a verb that is used 100 times less frequently regularizes 10 times as fast.

Different language

These findings completely underwhelmed most historical linguists. They already knew that commonly used words change more slowly, and the fact that some aspects of this trend could be quantified did not really interest them. "I don't think the numbers are very exciting," says Campbell. "I would much rather it be relativized to 'in general, more frequent words change more slowly'."

One reason that many linguists have been unreceptive to such work is that they are not trained in statistics, and are unsure of how to compare and evaluate this type of numerical model themselves. "They feel they've been asked to just accept things," says Tandy Warnow, a computer scientist at the University of Texas at Austin who works with linguists. And even those people, such as Warnow, who can evaluate the models say that they are too unsophisticated at this stage to pronounce firm dates or quantitative rules. She says that the biological models need to be tailored to language and that they should incorporate the sound and grammar changes that are so important to linguists.

"I don't think the numbers are very exciting."
— Lyle Campbell

Better representations

Paul Heggarty, a linguist at the McDonald Institute for Archaeological Research at the University of Cambridge, UK, is already trying to refine his models in this way. Heggarty is building network diagrams rather than trees to show how similar languages interrelate. He thinks that these can provide a better representation of the relationship between two languages when two cultures rub shoulders very closely and borrow words freely. Then the links between branches of the tree — equivalent to horizontal gene transfer — become more important than the vertical branching. "It is entirely natural for languages to stand in complex cross-cutting relationships to each other that may not be compatible with any branching genealogy at all," he says.

As part of the network building, Heggarty is also trying to assign more subtle values to word changes than zeroes and ones. A superficial analysis of the word 'dog', for example, might show that the English word is not cognate to the German word (*hund*) and score 1, or 'changed' for the pair. But if the English word 'hound' is chosen instead, it creates a match and would score 0. Because 'hound' isn't the main word for 'dog' in English, Heggarty would score it somewhere in between 0 and 1, perhaps 0.4. He hopes that this type of refined method can create networks that reproduce the real relationships between very closely related languages and, by extension, reveal something about the histories of the peoples who spoke them in the past.

Getting quantititive

Such model tweaking is unlikely to win over the historical linguists, but at least some are beginning to warm to the methods. Campbell acknowledges that the sheer number-crunching power of computer models can speed up the good old comparative method. And he sees the appeal in getting a bit more quantitative. If the field does not become more statistical and accountable, he points out, it may lose respect by those in other disciplines. "I think we'd like the legitimacy," he says.

Another fan is Harvard University's Steven Pinker, who famously appreciates language in all its fullness. "There has got to be information in the statistics of language overlap that you simply can't exploit by looking at it intuitively, by eyeballing," he says. "Linguists have been slow in accepting that extra dollop of information that statistics provides, even if there are errors, even if there is noise."

Noise — of the statistical kind — is not comfortable territory for many historical linguists when precious words such as *khun* are at stake. So perhaps the onus now lies on the newcomers to show that their methods will not drown out languages, or their rich and idiosyncratic narrative. "Hope," Pinker says, "is not that the older generation of linguists will lay down their arms; hope is that the younger generation will follow their noses to what is fruitful." ■

Emma Marris writes for *Nature* from Columbia, Missouri.

1. Dyen, I., Kruskal, J. B. & Black, P. *Trans. Am. Phil. Soc.* **82,** 1–132 (1992).
2. Pagel, M. *et al. Nature* **449,** 717–720 (2007).
3. Gray, R. D. & Atkinson, Q. *Nature* **426,** 435–439 (2003).
4. Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J. & Pagel, M. *Science* **319,** 588 (2008).
5. Lieberman, E. *et al. Nature* **449,** 713–716 (2007).