

Automated “time machine” reconstructs ancient languages

Yasmin Anwar, UC Berkeley



Ancient languages hold a treasure trove of information about the culture, politics and commerce of millennia past. Yet, reconstructing them to reveal clues into human history can require decades of painstaking work. Now, scientists at the University of California, Berkeley, have created an automated “time machine,” of sorts, that will greatly accelerate and improve the process of reconstructing hundreds of ancestral languages.

In a compelling example of how “big data” and machine learning are beginning to make a significant impact on all facets of knowledge, researchers from UC Berkeley and the University of British Columbia have created a computer program that can rapidly reconstruct “proto-languages”—the linguistic ancestors from which all modern languages have evolved. These earliest-known languages include Proto-Indo-European, Proto-Afroasiatic and, in this case, Proto-Austronesian, which gave rise to languages spoken in Southeast Asia, parts of continental Asia, Australasia and the Pacific.

“What excites me about this system is that it takes so many of the great ideas that linguists have had about historical reconstruction, and it automates them at a new scale: more data, more words, more languages, but less time,” said Dan Klein, an associate professor of computer science at UC Berkeley and co-author of the paper published online today (Feb. 11) in the journal *Proceedings of the National Academy of Sciences*.

The research team’s computational model uses probabilistic reasoning—which explores logic and statistics to predict an outcome—to reconstruct more than 600 Proto-Austronesian languages from an existing database of more than 140,000 words, replicating with 85% accuracy what linguists had done manually. While manual reconstruction is a meticulous process that can take years, this system can

Automated “time machine” reconstructs ancient languages

Published on Research & Development (<http://www.rdmag.com>)

perform a large-scale reconstruction in a matter of days or even hours, researchers said.

Not only will this program speed up the ability of linguists to rebuild the world’s proto-languages on a large scale, boosting our understanding of ancient civilizations based on their vocabularies, but it can also provide clues to how languages might change years from now.

“Our statistical model can be used to answer scientific questions about languages over time, not only to make inferences about the past, but also to extrapolate how language might change in the future,” said Tom Griffiths, associate professor of psychology, director of UC Berkeley’s Computational Cognitive Science Lab and another co-author of the paper.

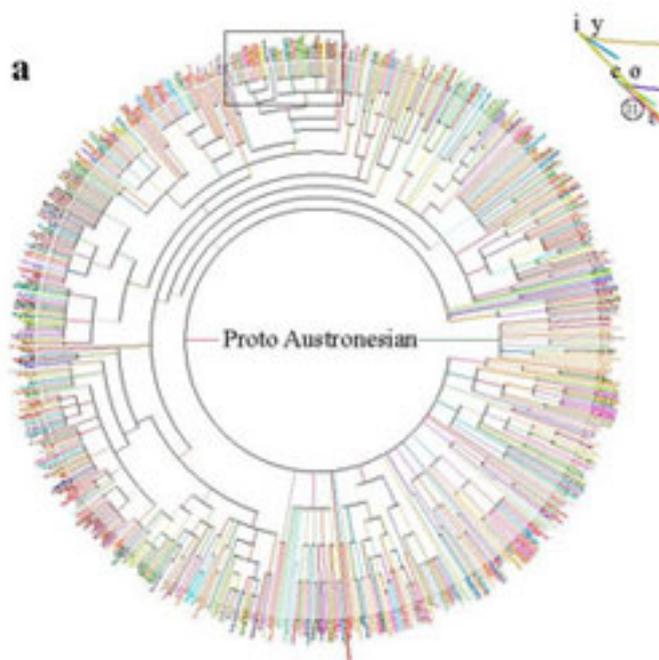
The discovery advances UC Berkeley’s mission to make sense of big data and to use new technology to document and maintain endangered languages as critical resources for preserving cultures and knowledge. For example, researchers plan to use the same computational model to reconstruct indigenous North American proto-languages.

Humans’ earliest written records date back less than 6,000 years, long after the advent of many proto-languages. While archeologists can catch direct glimpses of ancient languages in written form, linguists typically use what is known as the “comparative method” to probe the past. This method establishes relationships between languages, identifying sounds that change with regularity over time to determine whether they share a common mother language.

"To understand how language changes—which sounds are more likely to change and what they will become—requires reconstructing and analyzing massive amounts of ancestral word forms, which is where automatic reconstructions play an important role," said Alexandre Bouchard-Côté, an assistant professor of statistics at the University of British Columbia and lead author of the study, which he started while a graduate student at UC Berkeley.

Automated “time machine” reconstructs ancient languages

Published on Research & Development (<http://www.rdmag.com>)



The UC Berkeley computational model is based on the established linguistic theory that words evolve along the branches of a family tree—much like a genealogical tree—reflecting linguistic relationships that evolve over time, with the roots and nodes representing proto-languages and the leaves representing modern languages.

Using an algorithm known as the Markov chain Monte Carlo sampler, the program sorted through sets of cognates, words in different languages that share a common sound, history and origin, to calculate the odds of which set is derived from which proto-language. At each step, it stored a hypothesized reconstruction for each cognate and each ancestral language.

“Because the sound changes and reconstructions are closely linked, our system uses them to repeatedly improve each other,” Klein said. “It first fixes its predicted sound changes and deduces better reconstructions of the ancient forms. It then fixes the reconstructions and re-analyzes the sound changes. These steps are repeated, and both predictions gradually improve as the underlying structure emerges over time.”

Source: [University of California, Berkeley](http://newscenter.berkeley.edu/2013/02/11/ancientlanguages/) [1]

Source URL (retrieved on 05/23/2013 - 9:25pm):

<http://www.rdmag.com/news/2013/02/automated-%E2%80%9Ctime-machine%E2%80%9D-reconstructs-ancient-languages>

Links:

[1] <http://newscenter.berkeley.edu/2013/02/11/ancientlanguages/>