COMMENTARY

The descent of words

Quentin D. Atkinson[1]
School of Psychology, University of Auckland, Auckland 1142, New Zealand

The word for "sky" in the indigenous Saaroa language of Taiwan is *laŋica*. Across the South China Sea in the Philippines, the speakers of Ilonggo use *laŋit*, whereas, on the far-flung islands of the Pacific, Hawaiians say *lani* and Rarotongans and New Zealand Maori *raŋi* (1). Systematic sound correspondences between many such words tell us that these languages have evolved from a common ancestor to form part of the Austronesian language family. By meticulously comparing the sounds of words across many languages, linguists can learn about the genealogical relationships between languages and the people who speak them, how sounds change through time and even how long-extinct ancestral languages would have sounded. In PNAS, Bouchard-Côté et al. (2) automate this process by using probabilistic models of sound change to trace the evolution of thousands of words across more than 600 Austronesian languages.

The conventional technique for studying language change on the basis of contemporary variation is known as the comparative method (3). This approach identifies shared "cognates" between putatively related languages. Cognates are homologous words of similar meaning that show systematic sound correspondences indicating common ancestry (Fig. 1). Since the 19th century, historical linguists have understood that sound changes occur in a regular but context-sensitive way across the vocabulary of a language. Hence, where Hawaiian has *lani* and *lima* for "sky" and "five," Rarotongan and Maori have *raŋi* and *rima*, reflecting a shift in their ancestral lineage from this *l* sound to *r*. In deciding whether two words are genuinely cognate, linguists can therefore look beyond superficial similarities by attempting to reconstruct a protolanguage (the common ancestor of the languages in question) and identify regular sound changes acting across the sound systems of its descendants.

The rigorous application of the comparative method can be a complex and labor-intensive task. Accurate comparisons between words must incorporate likely changes to pronunciation and the phonological system and correctly align words allowing for insertions, deletions, metathesis (reversals, such as Old English *brid* to the modern *bird*), reduplication (such as Maori *paki* "to pat" vs. *pakipaki* "to clap"), and haplology (loss of repeated syllables, such as English *library* vs. the colloquial *libry*) among numerous other kinds of change. Change can also be context dependent. For example, in Proto-Germanic, stops (**p*, **t*, and **k*) became voiced (**b*, **d*, and **g*) but only after an unstressed syllable (Verner's Law); in other contexts, a different rule applied. This predictability allows linguists to distinguish true cognates from chance resemblances (such as the word for "eye" in Maori, *mata*, and Greek, *mati*) or likely borrowings (e.g., English *mountain* borrowed from Old French *montaigne*). All this is done at the same time as evaluating the underlying ancestral genealogy, which depends on and informs the observed patterns of sound change. The result is an iterative process in which multiple parameters are being optimized simultaneously across hundreds or thousands of data points.

Evolutionary biologists face an analogous and equally complex task in reconstructing species ancestry from gene sequence data (4). Like historical linguists, they seek to simultaneously infer homology, ancestral states, the ancestral genealogy, and underlying models of change. Biologists must also deal with alignment problems (including insertions, deletions, reversals, and reduplications) (5), context-dependent rates of change (6), multiple data types (7), and horizontal transmission (8). In response to these challenges, biologists have developed a suite of computational modeling tools that can efficiently explore parameter space and quantify uncertainty for even complex models and large datasets.

Recently, these tools have been applied to language data to model the evolution of words through time and test hypotheses about the origins of major language families (9–11). Until now, most computational models of vocabulary evolution have ignored information on the sounds of specific words, preferring simpler models of the gain and loss of cognates through time. However, this relies on existing cognate judgments from expert linguists, discards useful information in the source data, and cannot provide insight into the process of sound change.

Bouchard-Côté et al. (2) bring evolutionary modeling and historical linguistics one step closer by developing a probabilistic model of sound change that automates the process of

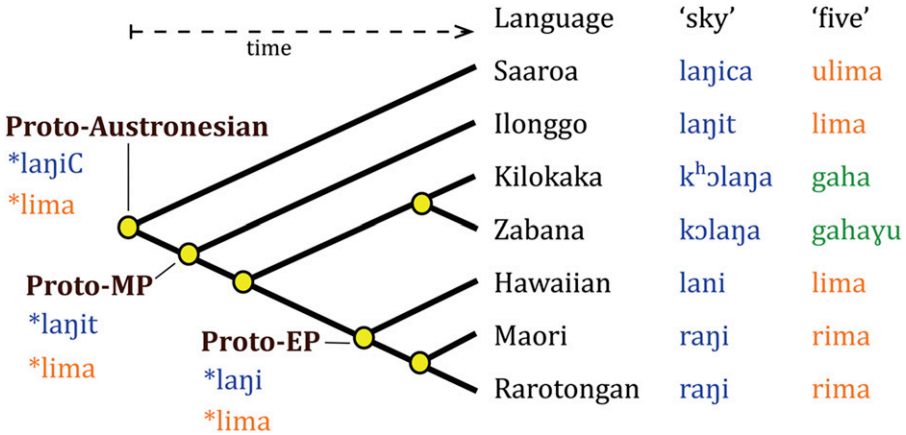

Fig. 1. Reconstructing the descent of words on a language tree. This example shows words for "sky" and "five" in a selection of seven modern Austronesian languages, together with the established language family tree connecting them (1). The three different cognate word forms are color-coded blue, orange, and green. Yellow dots at the nodes of the tree represent ancestral protolanguages. Reconstructed protoforms for the two larger cognate sets (blue and orange) are shown for Proto-Austronesian (the base of the tree), Proto-Malayo-Polynesian (Proto-MP), and Proto-Eastern-Polynesian (Proto-EP).

Author contributions: Q.D.A. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 4224.

[1]E-mail: q.atkinson@auckland.ac.nz.

ancestral state reconstruction and cognate assignment directly from vocabulary data. Previous attempts to solve this problem have been restricted to small datasets (12, 13), limiting the power and utility of the methods. Others have sought to quantify language diversification by using simple edit distances (14), but these efforts lack any explicit model of change or the ability to infer ancestral forms or cognates.

Bouchard-Côté et al.'s (2) approach adapts probabilistic string transducer algorithms developed by biologists for ancestral genome reconstruction and alignment (15). These computationally efficient algorithms make it possible to analyze large datasets and can handle many of the complexities of sound change considered by the comparative method. Bouchard-Côté et al. (2) infer ancestral sounds by estimating the probability of all possible sound changes occurring along each branch of the language family tree. By linking these probabilities across cognate sets, they can incorporate the regularity of sound change. The string transducer also allows for insertions, deletions, and a degree of context dependence. By adding into the model the further possibility of wholesale replacement with noncognate word forms, the method can reconstruct the birth and death of new cognates and so infer cognate words.

Based on two alternative Austronesian language trees (9, 16), Bouchard-Côté et al. (2) are able to reconstruct ancestral "proto-forms" for each cognate set. They benchmark their reconstructions against manual reconstructions of Proto-Oceanic (the common ancestor of modern languages from the Oceanic subgroup) and find an error rate midway between that achieved by randomly assigning cognate words from modern Oceanic languages and the level of disagreement between two linguists' manual reconstructions. Bouchard-Côté et al. (2) also compare cognate sets inferred under their model to known Oceanic cognate sets (1) and find they can group more than 90% of the words correctly.

One major limitation of the current implementation of Bouchard-Côté et al.'s (2) method is that it requires an existing language tree and so can only be applied to well-studied families in which the hard work of establishing the genealogy has already been done. In principle, however, the approach could be extended to simultaneously infer cognates and the tree directly from word string data. An analogous problem has already been solved in biology with the simultaneous estimation of gene alignment and phylogeny (5).

Regardless, by explicitly modeling probabilities of change across the tree, this new approach makes it possible to statistically test hypotheses that embody long-standing questions about the nature of sound change. Bouchard-Côté et al. (2) demonstrate this ability by revealing decisive support for the "functional load" hypothesis (17): the more work a sound contrast does in differentiating between words in a language, the less likely that contrast is to be lost. Identifying this pattern required integrating over thousands of data points and would simply not have been practical via manual reconstruction. The same tools could be used to answer questions about other functional dependencies and frequency effects (18), conditioning (19), and whether proposed laws are universal or family-specific.

Bouchard-Côté et al.'s contribution can be seen as a first step toward a comprehensive computational model of sound change.

Bouchard-Côté et al. (2) are careful to point out the limits of their current model and that it is not a replacement for careful linguistic scholarship. Besides not yet inferring the tree, the method falls short of being able to recover ancestral forms with the reliability of an expert linguist. Much of the shortfall may result from the fact that the string transducer algorithm does not permit metatheses, reduplications, or haplologies, and allows context dependency based only on the previous character in the string. However, these should be viewed as challenges to be solved, rather than inherent weaknesses of a computational approach. It is worth noting that biologists have achieved considerable success by starting with very simple models of complex phenomena and gradually increasing realism. Bouchard-Côté et al.'s (2) contribution can be seen as a first step toward a comprehensive computational model of sound change. Indeed, compared with the rudimentary models of nucleotide substitution first used by biologists, Bouchard-Côté et al.'s model of sound change is highly sophisticated. It seems reasonable to expect that computer algorithms will become an increasingly important tool for studying the descent of words. Although they cannot yet outcompete the grand masters of historical linguistics, Bouchard-Côté et al. show that they can certainly play the game.

1 Greenhill SJ, Blust R, Gray RD (2003–2013) Austronesian Basic Vocabulary Database. Available at http://language.psy.auckland.ac.nz/austronesian, accessed January 12, 2013.
2 Bouchard-Côté A, Hall D, Griffiths TL, Klein D (2013) Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc Natl Acad Sci USA* 110:4224–4229.
3 Campbell L, Poser WJ (2008) *Language Classification: History and Method* (Cambridge Univ Press, Cambridge, UK).
4 Atkinson QD, Gray RD (2005) Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Syst Biol* 54(4):513–526.
5 Suchard MA, Redelings BD (2006) BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22(16): 2047–2048.
6 Nevarez PA, DeBoever CM, Freeland BJ, Quitt MA, Bush EC (2010) Context dependent substitution biases vary within the human genome. *BMC Bioinformatics* 11(1):462.
7 Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. *Syst Biol* 53(1): 47–67.
8 Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104(3):870–875.
9 Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913):479–483.
10 Kitchen A, Ehret C, Assefa S, Mulligan CJ (2009) Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc Biol Sci* 276(1668):2703–2710.
11 Bouckaert R, et al. (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–960.
12 Ellison TM (2007) Bayesian identification of cognates and correspondences. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology* (Association for Computational Linguistics, Stroudsburg, PA), pp 15–22.
13 Oakes MP (2000) Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *J Quant Linguist* 7(3):233–243.
14 Brown CH, Holman EW, Wichmann S, Velupillai V (2008) Automated classification of the world's languages: A description of the method and preliminary results. *STUF Language Typology Universals* 61(4):285–308.
15 Holmes I, Bruno WJ (2001) Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* 17(9):803–820.
16 Lewis PM (2009) *Ethnologue: Languages of the World* (SIL, Dallas), 16th ed.
17 King R (1967) Functional load and sound change. *Language* 43:831–852.
18 Bybee JL (2001) *Phonology and Language Use* (Cambridge Univ Press, Cambridge UK).
19 Blust R (2004) *t to k: An Austronesian sound change revisited. *Oceanic Linguistics* 43(2):365–410.