John Benjamins Publishing Company

*Language Classification by Numbers.* By April McMahon and Robert McMahon. Oxford: Oxford University Press, 2005. Pp xvii, 265.

**Reviewed by Quentin D. Atkinson (University of Oxford)**

Today, evolutionary biology is about numbers…lots of numbers. Genes are the currency of modern biology and computers are employed to do the accounting. Computational phylogenetic methods incorporating statistical models of evolution are now routinely used to infer species' ancestry and test evolutionary hypotheses. This revolution was not driven by a sudden outbreak of numero — or techno — philia, but by the realization that a quantitative approach can provide powerful tools for analysing large amounts of data. By employing a rigorous statistical framework, uncertainty in results can be quantified and competing hypotheses evaluated. In addition, explicitly modelling the process of evolution means that any assumptions are made clear and their validity is testable. Using these methods, biologists have been able to answer an impressive range of questions about species evolution, including where and when certain species evolved, how they are related, what traits were present in ancestral populations and even what selective forces have operated in the past.

This quantitative revolution in evolutionary biology should be of special interest to historical linguists. Like species, languages are a product of evolution. There are, of course, some non-trivial differences between linguistic and biological evolution. For example, languages change much faster and horizontal transfer of hereditary material is probably more common between languages than species, at least within the animal kingdom. However, these differences are not fundamental. The words, phonemes, syntax and morphology of languages, are, like genes, passed on from generation to generation via a process of descent with modification. This common general mechanism means that evolutionary biologists and historical linguists are interested in similar questions and can use similar methods to answer these questions.

Over the last decade, there has been increasing interest in the application of computational phylogenetic methods from biology to the study of language change. These techniques are the quantitative equivalent of historical linguistics' comparative method. They infer the ancestral genealogy of a group of languages by combining the observed distribution of cognate words, phonemes and/or grammatical features with a set of assumptions about the process of language transmission and diversification that gave rise to those features. The fundamental

*Diachronica* 26:1 (2009), **125–133**. DOI 10.1075/dia.26.1.06atk
ISSN 0176–4225 / E-ISSN 1569–9714 © John Benjamins Publishing Company

difference in the new methods lies in their use of mathematical models of the evolutionary process and computerized search algorithms designed to evaluate competing solutions according to explicit optimality criteria.[1] This approach has been used to answer questions about the classification and expansion of major language families (Dunn et al. 2005, Forster & Toth 2003, Gray & Jordan 2000, Gray & Atkinson 2003, Holden 2002, McMahon & McMahon 2003, Rexova et al. 2003, Ringe et al. 2002), as well as to investigate patterns of lexical borrowing (Bryant et al. 2005) and factors affecting rates of language evolution (Atkinson et al. 2008, Pagel et al. 2007). The growing list of publications in the area has prompted a number of intersdisciplinary review papers (Gray et al. 2007, Marris 2008, Mesoudi et al. 2006, Whitfield 2008) to propose that historical linguistics (and the study of cultural evolution more generally) is on the verge of a quantitative revolution similar to that which has occurred in evolutionary biology.

The time is ripe then for *Language Classification by Numbers* (henceforth *LCBN*). M & M — a linguist and a population geneticist — present in their book a summary of major areas of quantitative research in historical linguistics together with their vision for the future of the field. Part introductory text and part manifesto, part linguistics and part biology, *LCBN* seeks not only to present the various quantitative methods that are available, but also to explain why these new methods might be useful and, presumably (the authors are coy about this), to see them used more widely. So, what can a curious reader learn? Will they be inspired to actually use the new methods? And what advances have occurred since the emergence of *LCBN* in 2005?

The opening chapters of *LCBN* are a helpful starting point for anyone interested in the application of quantitative phylogenetic methods to language change. The material constitutes a well-argued case for adopting a quantitative approach and provides a good introduction to some of the major issues in the field. This includes an excellent discussion of the pitfalls of working with comparative word lists, covering problems with onomatopoeia, list length, choosing (or failing to choose) suitable meaning boundaries and identifying borrowed words and chance resemblances. There are also straight-forward explanations of the rationale behind most of the principle methods of data coding and tree-building, as well as how network methods (Huson & Bryant 2004), which do not assume a strictly tree-like model, can be used to investigate patterns of borrowing between languages (Bryant et al. 2005). Although the breadth of coverage is impressive, the level of detail *LCBN* offers does vary between methods. A large portion of the material focusses

[1] The latter is particularly important in phylogenetic reconstruction where the number of possible genealogies is extremely large, even for small sample sizes. For ten languages, for example, there are over 34 million possible genealogical trees.

on lexicostatistics (Swadesh 1952, Swadesh 1955), an approach to tree-building based on similarity scores between languages. These methods throw away precious phylogenetic information by reducing comparative data to single pair-wise distance scores between languages and are especially sensitive to the effects of rate variation and borrowing (Blust 2000). Whilst interesting in terms of the history of the discipline, distance-based methods are being replaced by more powerful techniques that explicitly model the evolution of language features on a phylogeny and so do not suffer from the problem of information loss. Somewhat disappointingly, these newer approaches, such as parsimony, likelihood modelling and Bayesian phylogenetic inference techniques, now the methods of choice in biology, are given relatively short shrift in *LCBN*.[2]

One of the strengths of the book lies in M & M's use of computer simulations and well-studied test cases to evaluate the reliability of the quantitative methods they employ. A common misconception among critics of quantitative methods in historical linguistics is that they must be applied blindly, as if one is feeding words into a black box and getting numbers out the other end. In fact, by having to quantify and model the process of language change, researchers must be explicit about the assumptions they make and are in a position to test those assumptions systematically. M & M's step-by-step approach to applying each method — testing its validity on simulated data, then on well known case-studies and finally, where possible, using what they have learned to say something new — is, of course, just good science. However, it also serves as a rebuttal to the 'black box' criticism, nicely illustrating the iterative process by which methodological issues can be resolved and new knowledge gained.

Unsurprisingly for a book written by a linguist and a geneticist, the chapter on correlating linguistic and genetic data is also handled adeptly. Languages are spoken by people, which makes historical linguistics as much about the history of peoples as it is about languages. The field has much to contribute to our understanding of human history as part of a 'new synthesis' with archaeology and population genetics. Here M & M are at their most forthright about the need for quantitative methods, warning that "If we are genuinely interested in interdisciplinary research, and do not supply numbers of our choosing, we cannot be surprised if archaeologists and geneticists attempt to provide their own" (pp. 125). The pair review some of the landmark studies linking genes to languages, including Cavalli-Sforza, et al.'s (1988) controversial comparison of global genetic and linguistic phylograms. They highlight important caveats to do with the type of genetic and linguistic data being compared, how populations are defined and sampled and

[2] For those wanting more technical explanation of these methods, Joe Felsenstein's *Inferring Phylogenies* (2004) is the most current and thorough treatment.

how similarity in linguistic and genetic variation is measured. M & M also raise the interesting point that some of the disagreement between linguistic and genetic diversity, at least at the regional level, may be due to the common practice of removing 'borrowings' from linguistic data, effectively suppressing any evidence for recent language contact. It is worth noting here that the literature linking genes and languages has focussed almost exclusively on the global or regional scale and associations arising from shared history. Understandably, M & M's summary reflects this focus. However, since the publication of *LCBN*, two exciting studies have emerged which highlight the potential for research a) at a much smaller and more detailed scale and b) examining potential functional associations between genes and language. Lansing et al. (2007) looked at the association between Austronesian vocabulary terms and paternally inherited genetic markers on the small island of Sumba in Indonesia. They found support for a model coupling genetic admixture and language shift linked to the colonization of the island by Austronesian farmers. This small-scale analysis provides novel insight into micro-level processes shaping gene-language coevolution. With regard to functional links between genes and language, Dediu & Ladd (2007) have demonstrated that tonal languages are more common in populations with a higher proportion of ASPM and Microcephalin genetic variants, which are known to be related to brain growth and development, suggesting that languages may evolve to 'fit' their human genetic landscape.

Whilst most of *LCBN* can be seen as a defense of quantitative methods, M & M attack attempts to derive dates from linguistic data, even going as far as to call for a moratorium on any language-based dating. This stance is somewhat puzzling given the pair's expressed desire to see historical linguists contribute to the 'new synthesis' by providing more quantitative output. Without an independent linguistic time scale, the feasibility of linking language relationships to evidence from archaeology and genetics is greatly reduced. M & M's scepticism is principally directed at glottochronology (Swadesh 1955), a method for dating language trees based on the percentage of cognate words shared between languages and the assumption of a 'glottoclock', or constant rate of lexical replacement through time. Whilst problems with glottochronology are widely recognized[3] and M & M are right to criticize the method, more sophisticated dating techniques are now available that do not suffer from the same limitations.

Gray & Atkinson's (2003) analysis of Indo-European lexical data systematically addresses the problems with glottochronology and shows how these can be overcome using phylogenetic divergence time estimation techniques from evolutionary biology. Most notably, their approach does not assume constant rates

[3] See Bergsland & Vogt (1962) and Campbell's (2004) thorough summary of the pitfalls of glottochronology.

of change between words or through time and is able to quantify phylogenetic and calibration uncertainty in estimated dates and rates.[4] M & M give a favourable treatment of Gray & Atkinson (2003), even suggesting that "if phylogenetic approaches to dating are to be adopted, they should follow this kind of pattern" (p. 198). However, when evaluating the feasibility of dating, they lump Gray & Atkinson's method together with other "essentially glottochronological" approaches (p. 190) for which they offer only a general, philosophical critique. Their critique boils down to two related points which I consider below. First, languages do not change at a constant rate. Whilst some natural processes, like nucleotide subsitution or radioactive decay, are predictable and regular enough to base date estimates on, M & M argue language change is not such a process. As a result, whilst we can infer the order of events based on branching patterns in a tree, it is not possible to infer absolute dates. Second, the pair argue that even if we did know the rate of language change, borrowing and uncertainty about underlying family tree relationships make dating impossible and, hence, we need to sort out the methods of inferring phylogenies before we move on to dating.

In response to the first point, it is simply not true that language change is inherently more complex or less predicatble than genetic evolution. Both are incredibly complex systems that can be influenced by various contingencies on a number of time scales. Regardless, as with genetic data, there is clearly chronological information in linguistic diversity. The relevant question is "How much?" I do not need a calculator or very much knowledge of the languages to say with confidence that Maori and Samoan did not separate yesterday, or that French split off from Italian more recently than the Bronze Age. If we want to say something more precise (and more interesting), like, for example, whether the Austronesian language family is 5,000 or 10,000 years old or whether Bantu languages could have expanded with the spread of farming in sub-Saharan Africa around 3000 years ago, the sensible application of quantitative dating techniques can replace guess work with principled, hypothetico-deductive reasoning. Of course, uncertainty in any estimates is crucial and the result may sometimes be that there is not enough chronological information in the data to help us decide. Whether useful date information can be extracted depends on a number of factors including the quality of the data, the availablity of information about rates of change, and the hypothesis being tested. However, this does not mean we should avoid applying quantitative dating methods to answer questions we have about language divergence times.

[4] A series of papers have been published following Gray & Atkinson (2003), which offer a more detailed discussion of the methods and their validity (Atkinson et al. 2005, Atkinson 2006, Nicholls & Gray 2006).

Regarding the second point, there is no reason to think that inferring relationships between languages is any easier than inferring language divergence times. Indeed the former is in general more computationally demanding. Nor does the accurate estimation of divergence times require prior knowledge of the underlying family tree. For example, a Bayesian approach to phylogenetic inference allows divergence times to be estimated across the distribution of plausible trees, thus accounting for uncertainty in the underlying phylogeny. A nice analogy is that it is possible to infer the approximate position of the trunk of a tree by observing the location of its leaves — one does not necessarily need to know the complex pattern of branching underneath.

The book's final chapter sees a shift in the scale of analysis from inter-population language classification to intra-population dialectology. M & M show how phylogenetic methods can be used to classify regional dialects by comparing phonetic variation across population sub-groups.[5] Since dialect boundaries tend to remain fluid, the interpretation of dialect trees is problematic in that they are not representative of the historical process that has given rise to the observed data. M & M recognize this and describe the use of network methods (which do not assume tree-like evolution) as a tool for visualising dialect variation and identifying patterns of reticulation. Whilst this is a significant improvement over strictly tree-based classification, dialectology may gain more from looking to the methods that biologists use to study within — rather than between — population variation. Population genetics is concerned with the emergence and propagation of competing genetic variants within a population. As noted in a recent review of Darwinian approaches to culture, this biological process has obvious parallels with the innovation and spread of cultural replicators, including dialectical variants in speaker populations (Mesoudi et al. 2006). As a result, the tools that biologists have developed to study mutation, selection, drift and gene flow may be more useful than phylogenetic methods for studying the process of language change within populations.[6]

[5] The authors advocate a word comparison metric that compares forms "through a template consisting of the appropriate proto-form" (p. 218) and are critical of research using edit distance (the number of insertions, deletions or substitutions required to transform one string into another) as a proxy for phonetic difference. John Nerbonne has addressed these criticisms in detail in an earlier review of LCBN (Nerbonne 2007), so I will not discuss them here. It is worth noting, though, that both approaches show parallels with methods of sequence alignment in biology (e.g., Clustal W. Thompson et al. 1997) that could be usefully applied to language.

[6] See Croft (2000) and Mufwene (2001) for detailed discussion of how Darwinian concepts from population genetics can be applied to the study of language change.

Perhaps the reason that these methods are not discussed in *LCBN* is that, as the title of the book suggests, M & M are principally concerned with language *classification*. However, one of the major advantages of a quantitative approach to language evolution is the opportunity to model and investigate the *process* of language change. In the book's introduction M & M are clear that they see classification as the first step towards this goal. As with their discussion of dating, classification is presented as laying the foundation on which everything else must be built. It is true that developing accurate language classification is important, but it should not — indeed cannot — preceed all study of the process of change. An understanding of process and classification are inextricably linked — the one building on the other — and progress can only be achieved by moving both forward simultaneously, using new methods of classification to study the process of change and new knowledge about process to study classification. Quantitative approaches to dialectology, then, have much to offer historical linguistics by prodiving models of how changes actually occur within populations. Process can also be studied beyond the population level. Since the publication of *LCBN,* quantitative phylogenetic methods have been applied to whole language families, representing thousands of years of language change, to show that frequency and part of speech can explain most of the variance in rates of vocabulary replacement across meanings (Pagel et al. 2007) and that vocabulary evolves in punctuational bursts associated with the emergence of new languages (Atkinson et al. 2008). Using a quantitative framework to study how different factors affect language evolution is crucial to the development of better methods of language classification and can help to shed light on the basic mechanisms underlying language change over long time-scales.

Despite recent advances in the field, LCBN is still a worthwhile starting point for anyone wanting to know about quantitative approaches to historical linguistics. Whether you are a linguist interested in quantitative methods or a mathematician interested in language, there is plenty of valuable insight and information on offer. It remains to be seen, however, whether the book is sufficiently inspiring to motivate more linguists to use quantitative methods. Historical linguistics is blessed with centuries of careful work that has produced a wealth of data and understanding, but many linguists seem reluctant to combine these resources with quantitative analysis tools. Perhaps the best motivator will come from looking to evolutionary biology and the realisation that if linguists don't do it, someone else will.

## References

Atkinson, Quentin D. 2006. "Are accurate dates an intractable problem for historical linguistics?" *Mapping our Ancestry: Phylogenetic methods in anthropology and prehistory* ed. by Carl Lipo, Michael O'Brien, Stephen Shennan & Mark Collard, 269–296. Chicago: Aldine.

Atkinson, Quentin D., Geoff K. Nicholls, David Welch & Russell D. Gray. 2005. "From Words to Dates: Water into wine, mathemagic or phylogenetic inference?" *Transactions of the Philological Society* 103.193–219.

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill & Mark Pagel. 2008. "Languages evolve in punctuational bursts". *Science* 319.588.

Bergsland, Knut & Hans Vogt. 1962. "On the validity of glottochronology". *Current Anthropology* 3.115–153.

Blust, Robert. 2000. "Why lexicostatistics doesn't work: The 'universal constant' hypothesis and the Austronesian languages". *Time Depth in Historical Linguistics* ed. by Colin Renfrew, April McMahon & Larry Trask, 311–332. Cambridge: The McDonald Institute for Archaeological Research.

Bryant, David, Flavia Filimon & Russell D. Gray. 2005. "Untangling our past: Languages, trees, splits and networks". *The Evolution of Cultural Diversity: Phylogenetic approaches* ed. by Ruth Mace, Clare J. Holden & Stephen Shennan, 69–85. London: UCL Press.

Campbell, Lyle. 2004. *Historical Linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Cavalli-Sforza, Luigi L., Alberto Piazza, Paolo Menozzi & Joanna Mountain. 1988. "Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data". *Proceedings of the National Academy of Sciences* 85.6002–6006.

Croft, William. 2000. *Explaining Language Change: An evolutionary approach*. Harlow, England: Longman.

Dediu, Dan & D. Robert Ladd. 2007. "Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin". *Proceedings of the National Academy of Sciences of the United States of America* 104.10944–10949.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. "Structural phylogenetics and the reconstruction of ancient language history". *Science* 309.2072–2075.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland: Sinauer Associates.

Forster, Peter & Alfred Toth. 2003. "Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European". *Proceedings of the National Academy of Sciences of the United States of America* 100.9079–9084.

Gray, Russell D. & Fiona M. Jordan. 2000. "Language trees support the express-train sequence of Austronesian expansion". *Nature* 405.1052–1055.

Gray, Russell D. & Quentin D. Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin". *Nature* 426.435–439.

Gray, Russell D., Simon J. Greenhill & Robert M. Ross. 2007. "The pleasures and perils of Darwinizing culture (with phylogenies)". *Biological Theory* 2.360–375.

Holden, Claire J. 2002. "Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis". *Proceedings Biological Sciences* 269.793–799.

Huson, Daniel & David Bryant. 2004. "Neighbor-Net: An agglomerative method for the construction of phylogenetic networks". *Molecular Biology and Evolution* 21.255–265.

Lansing, J. Stephen, Murray P. Cox, Sean S. Downey, Brandon M. Gabler, Brian Hallmark, Tatiana M. Karafet, Peter Norquest, John W. Schoenfelder, Herawati Sudoyo, Joseph C. Watkins & Michael F. Hammer. 2007. "Coevolution of languages and genes on the island of Sumba, Eastern Indonesia". *Proceedings of the National Academy of Sciences of the United States of America* 104.16022–16026.

Marris, Emma. 2008. "The language barrier". *Nature* 453.446–448.

McMahon, April & Robert McMahon. 2003. "Finding families: Quantitative methods in language classification". *Transactions of the Philological Society* 101.7–55.

Mesoudi, Alex, Andrew Whiten & Kevin N. Laland. 2006. "Towards a unified science of cultural evolution". *Behavioral and Brain Sciences* 29.329–383.

Mufwene, Salikoko. 2001. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Nerbonne, John. 2007. "Review of April McMahon & Robert McMahon Language Classification by the Numbers". *Linguistic Typology* 11.425–436.

Nicholls, Geoff K. & Russell D. Gray. 2006. "Quantifying uncertainty in a stochastic dollo model of vocabulary evolution". *Phylogenetic Methods and the Prehistory of Languages* ed. by James Clackson, Peter Forster & Colin Renfrew, 161–172. Cambridge: The McDonald Institute for Archaeological Research.

Pagel, Mark, Quentin D. Atkinson & Andrew Meade. 2007. "Frequency of word-use predicts rates of lexical evolution throughout Indo-European history". *Nature* 449.717–720.

Rexová, Katerina, Daniel Frynta & Jan Zrzavy. 2003. "Cladistic analysis of languages: Indo-European classification based on lexicostatistical data". *Cladistics-the International Journal of the Willi Hennig Society* 19.120–127.

Ringe, Donald, Tandy Warnow & Ann Taylor. 2002. "Indo-European and computational cladistics". *Transactions of the Philological Society* 100.59–129.

Swadesh, Morris. 1952. "Lexico-statistic dating of prehistoric ethnic contacts". *Proceedings of the American Philosophical Society* 96.453–463.

Swadesh, Moris. 1955. "Towards greater accuracy in lexicostatistic dating". *Journal of American Linguistics* 21.121–137.

Thompson, Julie D., Toby J. Gibson, Frédéric Plewniak, François Jeanmougin & Desmond G. Higgins. 1997. "The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools". *Nucleic Acids Research* 25.4876–4882.

Whitfield, John. 2008. "Across the curious parallel of language and species evolution". *PLoS Biology* 6.1370–1372.

## *Reviewer's address*

Quentin D. Atkinson
Institute of Cognitive and Evolutionary Anthropology
University of Oxford
64 Banbury Rd
Oxford OX2 6PN, United Kingdom

quentin.atkinson@anthro.ox.ac.uk