

Review Article

Robert Blust

Let's talk about trees: Genetic relationships of languages and their phylogenetic representation

Edited by RITSUKO KIKUSAWA and LAWRENCE A. REID

Osaka: Senri Ethnological Studies 98, National Museum of Ethnology.
2018. Pp. 176. Maps, Plates, Tables, Figures.

This volume contains an introduction and eight papers presented at an international symposium 'Let's Talk about Trees', which was organised by Ritsuko Kikusawa and hosted by the National Museum of Ethnology of Osaka, Japan, in February 2013. The stated purpose of the meeting was to evaluate the pros and cons of the classic tree model of historical linguistics in describing the order of splits within a language family. Because the problem of modelling relationships of descent is common to other disciplines, contributors were invited from a range of academic disciplines, including not only linguistics, but also what is described on page one as 'cladistics', 'biology' and 'genetics', although cladistics is clearly a part of biological taxonomy, and not an independent discipline.

Kikusawa and Lawrence Reid note in their Introduction to this edited volume that the tree model in linguistics was proposed by the German Indo-Europeanist August Schleicher in 1853, and challenged by his former student Johannes Schmidt in 1872, in particular with regard to Schleicher's recognition of a 'Germano-Slavic' branch of Indo-European, which reflects a history of borrowing among contiguous speech communities rather than an exclusive period of historical development for the Germanic and Slavic languages. In this and other ways, the reader is reminded that the family tree is an idealisation of the actual process of language split which, among other things, assumes abrupt and complete separation of language communities and isolation from external contacts, although this is rarely attested in observable cases.

Chapter 2, 'Tree and network in systematics, stemmatics, and linguistics: Structural model selection in phylogeny reconstructions' by Nobuhiro Minaka, a specialist in evolutionary biology and biostatistics, argues that researchers in different historical disciplines, including evolutionary biology, textual stemmatics, and historical linguistics, have independently struck upon a common set of principles for inferring phylogenetic relationships. Minaka argues that these disciplines have a common

Robert Blust is Professor at the Department of Linguistics, University of Hawai'i. Correspondence in connection with this article may be addressed to: blust@hawaii.edu.

historical core, where ‘history’ is defined as ‘spatiotemporal modification’, and that this opposes them to physics and chemistry, which do not. While this contrast might be generally accepted, it can be challenged, since cosmologists commonly believe that the laws of physics as we know them evolved from an earlier stage in which they did not apply, and that they consequently also have a ‘history’.¹ Much of Minaka’s chapter is concerned with a discussion of the ‘cladistics wars’ of the 1960s and 1970s, which pitted proponents of the view that a cladistic classification must have a historical interpretation vs those who believed it could be used for general classificatory purposes. In many ways this seems like a side issue, since the central debate in biological taxonomy has always been between those who insist that an accurate reflection of history is only possible with cladistics (strict trees), and is obscured by phenetics (strict similarity); for a concise and balanced review of these issues see Stephen Jay Gould’s essay, ‘What, if anything, is a zebra?’² Minaka’s discussion here could have benefited from stating that the cladistics/phenetics divide in biological taxonomy corresponds closely to subgrouping by exclusively shared innovations vs lexicostatistics (subgrouping by percentage of shared cognates) in historical linguistics.

Chapter 3, ‘Inferring population phylogeny from genetic data’ by Ryosuke Kimura, a specialist in evolutionary molecular genetics, is concerned exclusively with the classification of biological populations by genetic markers, and so lacks a direct connection with the title of the volume. Applications to linguistics are mentioned only once or twice very briefly in passing. Historical linguists are sure to pause at the statement (p. 25) that ‘The discipline of phylogenetics has been developed mainly in studies of taxonomy and molecular evolution, and it has been applied to other academic fields, such as linguistics.’ In fact, the distinction between innovation and retention (synapomorphy and symplesiomorphy), central to cladistic classifications and hence to biological phylogenetics, was first stated explicitly by Karl Brugmann in the late nineteenth century, and was widely used in linguistics for decades before the same principle was proposed in biology by Willi Hennig in the mid-twentieth century.³

Chapter 4, ‘Jackknifing the black sheep: ASJP classification performance and Austronesian’, by Søren Wichmann and Taraka Rama (W&R) addresses problems in Austronesian (AN) linguistic classification. Eric Holman et al.⁴ proposed 40 basic lexical items from which a database of translation equivalents has since been compiled ‘in the majority of the world’s languages’ (p. 39). This database was then used to develop an Automated Similarity Judgment Program (ASJP). According to the authors ‘Since 2008, the preferred approach to computing distances among languages for further input to various analyses has been a modified version of the

1 Steven Weinberg, *The first three minutes: A modern view of the origin of the universe* (New York: Basic Books, 2015 [1977]).

2 Stephen Jay Gould, *Hen’s teeth and horse’s toes: Further reflections in natural history* (New York: W.W. Norton, 1984), pp. 355–65.

3 Karl Brugmann, ‘Zur Frage nach den Verwandtschaftsverhältnissen der indogermanischen Sprachen’, *Internationale Zeitschrift für allgemeine Sprachwissenschaft* (1884): 226–56; Willi Hennig, *Phylogenetic systematics* (Urbana: University of Illinois Press, 1966 [1950]).

4 Eric W. Holman, Søren Wichmann, Cecil H. Brown, Viveka Velupillai, Andre Müller and Dik Bakker, ‘Explorations in automated language classification’, *Folia Linguistica* 42, 3–4 (2008): 331–54.

Levenshtein or “edit” distance, called LDND.’ The reader is told that this method of constructing phylogenetic trees has nearly always yielded results that agree closely with those reached by specialists who work with the language families in question. The major exception is AN, where some proposed branches as reflected in the ETHNOLOGUE (‘black sheep’) are not supported by the ASJP.

The first thing to note in commenting on this issue is that distance-based methods of linguistic classification are notoriously inaccurate as indicators of history. The classic example, which held its own for several decades before collapsing under the weight of massive counterevidence, is lexicostatistics, which may yield valid results in particular cases, but is equally likely to yield erroneous results.⁵ The reason that lexicostatistics fails to consistently give results that agree with history as inferred by qualitative methods in linguistics or by archaeology, is that it does not distinguish innovations from retentions. Consequently, it often yields groups that are formed on the basis of common retentions. Where lexical retention rates are uniform or nearly so over whole populations of languages little distortion is introduced, but where there is significant variation in retention rate, historical inferences based on lexicostatistics can go wildly wrong.⁶ Biologists interested in phylogenetic classification will recognise this as the problem that cladists see with phenetics.

It is surprising that the lesson of lexicostatistics in showing the pitfalls of classifications using distance-based measures has not been learned. The LDND is simply another variant of this conceptually flawed procedure.⁷ This can be illustrated with an artificial (but realistic) data sample such as the following:

Initial state *pakut
 Language A pakut
 Language B pakut
 Language C fahu?
 Language D hou

Using standard Levenshtein edit distances languages A and B are clearly closest, since the edit distance is zero. Language C differs from each of these by three edits, $p : f$, $k : h$ and $t : ?$. Language D differs from each of the other languages by four edits. Languages C and D probably subgroup with one another as a result of the lenition of voiceless stops, but this is hardly apparent, since the use of distance measures

5 Robert Blust, ‘Why lexicostatistics doesn’t work: The “universal constant” hypothesis and the Austronesian languages’, in *Time depth in historical linguistics*, ed. Colin Renfrew, April McMahon and Larry Trask (Cambridge: McDonald Institute for Archaeological Research, 2000), pp. 311–31.

6 Isidore Dyen, *A lexicostatistical classification of the Austronesian languages*, Indiana University Publications in Anthropology and Linguistics, and Memoir 19 of the International Journal of American Linguistics (Baltimore: The Waverly Press, 1965); George W. Grace, ‘Austronesian lexicostatistical classification: A review article’, *Oceanic Linguistics* 5 (1966): 13–31; Robert Blust, ‘The Austronesian homeland: A linguistic perspective’, *Asian Perspectives* 26, 1 (1984/1985): 45–67; Blust, ‘Why lexicostatistics doesn’t work’.

7 Simon J. Greenhill, ‘Levenshtein distances fail to identify language relationships accurately’, *Computational Linguistics* 37 (2011): 689–98.

unites Languages A and B on the basis of retentions, and shows Language C as being closer to Languages A and B than it is to Language D.

W&R attempt to improve the performance of standard Levenshtein distances by processes they call ‘normalisation’ and division (hence LDND). Despite these efforts the basic flaw remains: the LDND lacks the critical property characteristic of cladistic classifications in biology and their equivalent in historical linguistics, namely a distinction between innovations and retentions. Examples of AN languages that fare poorly using such distance measures are Sa’ban with other Dayic languages,⁸ Segai or Modang with their close relatives Kayan and Murik,⁹ Tsat with almost any other AN language, including its close relatives within Chamic,¹⁰ Tsou, or Palauan with almost any other AN language,¹¹ or any language at all that has undergone extensive sound change over a relatively short time, as this will inflate Levenshtein distances in relation to actual separation times.

Part of the problem with the W&R procedure is the issue whether a tree structure is the only appropriate means for representing language splits. This arises with Central-Malayo-Polynesian (CMP).¹² Whereas some phonological innovations are shared across many CMP languages, few if any of these include the entire group, suggesting that CMP began as a dialect chain due to rapid population movement, with subsequent diffusion of innovations among communities that were closely related at the time. Since this group is not readily represented in a dendrogram or tree structure the authors essentially withdraw from evaluating it, holding that ‘discrepancies between ASJP and ETHNOLOGUE with regard to the classification of CMP languages cannot be used as an argument of the inadequacy of ASJP’.

The other major Austronesian subgroup that is addressed is South Halmahera-West New Guinea (SHWNG), and here there is an explicit conflict, as SH languages such as Buli, Sawai, Giman or East Makian are not subgrouped with WNG languages such as Numfor, Biak, Dusner or Mor, despite persuasive evidence that all of these languages (and others) were once part of a dialect chain defined by (1) loss of final vowels, and (2) loss of *k, where (1) began in the far west and spread east without reaching the end of the chain, while (2) began in the far east and spread west in a similar fashion. What is most noteworthy about these innovations is that they overlapped in the middle of the chain, so that in Buli (nearer the western end) the order was 1, 2, while in Numfor (nearer the eastern end) it was 2, 1, a

8 Robert Blust, ‘Language, dialect and riotous sound change: The case of Sa’ban’, in *Papers from the Ninth Annual Meeting of the Southeast Asian Linguistics Society, 1999*, ed. Graham W. Thurgood (Tempe: Arizona State University, Program for Southeast Asian Studies, 2001), pp. 249–359.

9 Alexander D. Smith, ‘The languages of Borneo: A comprehensive classification’ (PhD diss., University of Hawai’i).

10 Graham Thurgood, *From ancient Cham to modern dialects: Two thousand years of language contact and change*, Oceanic Linguistics Special Publication No. 8 (Honolulu: University of Hawai’i Press, 1999).

11 Shigeru Tsuchida, *Reconstruction of Proto-Tsouic phonology*, Study of Languages and Cultures of Asia and Africa Monograph Series No. 5 (Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, 1976); Robert Blust, ‘Palauan historical phonology: Whence the intrusive velar nasal?’, *Oceanic Linguistics* 48 (2009): 307–36.

12 Robert Blust, ‘Central and Central-Eastern Malayo-Polynesian’, *Oceanic Linguistics* 32 (1993): 241–93.

clear indication of an earlier chaining situation.¹³ This relationship is completely lost in the ASJP approach.

Another problem with the ASJP results is that many languages that W&R name are virtually unknown, making debate about classification difficult. Space does not permit full discussion of their procedures, but W&R argue that if the fit between ASJP results and those of the standard view is improved by removing a language (a procedure called ‘jackknifing’) then that language is inaccurately classified by ASJP. Even with this qualification the authors are unconvinced of the reality of SHWNG, but given the evidence of shared phonological innovations discussed above what is at issue should be its composition, not its validity.

Chapter 5, ‘Freeing the comparative method from the tree model: A framework for historical glottometry’, by Siva Kalyan and Alexandre François, raises a question that has been asked at least since the advent of Wave Theory in 1872, namely is the tree model the only useful method of representing language splits? It thus addresses issues just raised in discussing the modelling of language change in the previous chapter.

The authors of this chapter remind the reader of facts that have long been known, namely that innovations are not always exclusively shared, but may be cross-cutting — that is, may support conflicting subgrouping inferences. The great value of this chapter is that it provides a novel and extremely helpful manner of representing the fact that innovations may be cross-cutting in a way that still allows clear visual representations of the history they represent. Intersection diagrams (called ‘glottometric diagrams’) show this not only by connecting a given language branch with more than one other branch, but most importantly by using the thickness of lines representing shared isoglosses to indicate the relative strength of these connections in strictly quantitative terms, determined by counting numbers of shared innovations supporting each connection. I personally find this one of the most original and valuable papers on language classification that I have read in many years. It is clearly written, with many useful diagrams, and its use of quantitative measures to indicate the relative strength of competing hypotheses is sure to become a tool that others adopt in future work.

The only shortcoming that I see in this chapter is that it may leave the misleading impression that the *Stammbaum*, or family tree diagram, is less useful than it really is. Despite its widely recognised idealisation of the actual process of language split, it has been commonly assumed since at least my student days that the family tree is a better approximation to reality for more distant relationships within a language family, and that the diffusion model works best for dialect networks and their descendants. There is little controversy, for example, about the reality of discrete Indo-European subgroups such as Germanic, Slavic or Indo-Iranian (Schleicher’s original ‘Slavo-Germanic branch’ notwithstanding!), or Austronesian subgroups such as Malayo-Polynesian, Oceanic or Polynesian. As more time passes language communities that were once united move further apart, reducing the chances of contact, and

13 Robert Blust, ‘Eastern Malayo-Polynesian: A subgrouping argument’, in *Second International Conference on Austronesian Linguistics: Proceedings*, ed. S.A. Wurm and Lois Carrington, *Fascicle 1*: 181–234 (Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University [ANU], 1978).

become more divergent, reducing the ease of communication and ready diffusion. In the Austronesian case the discreteness of subgroups also appears to correlate with significant pauses in a millennia-long expansion out of Taiwan to the farthest reaches of the Pacific: where there was rapid movement language splits do not reflect a tree-like structure, but where there were pauses they do.

Chapter 6, 'Modeling the linguistic situation in the Philippines', by Lawrence A. Reid looks at the languages of the Philippine archipelago in relation to problems of classification and the representation of descent. This chapter focuses mainly on two issues: first, how to represent the descent of Philippine languages spoken by Negrito populations, which seem clearly to have undergone language shift at some fairly remote period,¹⁴ and second, whether there is a Philippine subgroup.

Whereas the first of these issues has garnered virtually universal support, the second has divided the scholarly community. Reid holds that no phonological innovations mark off a Philippine subgroup, a strike against supporters of Proto-Philippines. However, this is not quite true, since the *d/z distinction, which is preserved in much of western Indonesia, various parts of eastern Indonesia, Chamorro, and Proto-Oceanic, has not been reported in any Philippine language. He acknowledges that numerous lexical distributions appear to be shared only by members of the proposed Philippine subgroup, but attributes these to 'extensive networking resulting from trade' (p. 101). This latter point merits at least a brief comment.

David Zorc presented 98 putative Proto-Philippine lexical innovations, 23 of which were classed as 'widespread', and 75 as 'selective' (found in only a few languages).¹⁵ In commenting on these, Malcolm Ross, whose position mirrors that of Reid, stated 'Ninety-eight lexical innovations is a substantial number'.¹⁶ At the same time he speculated whether 'some of Zorc's vocabulary items have been retained from Proto Malayo-Polynesian but lost in extra-Philippine Malayo-Polynesian languages,' since these latter may be 'descended from only a very small number of speech communities emanating out of the southern Philippines'. Although Zorc now questions the validity of Proto-Philippines,¹⁷ the number of putative Philippine-only cognate sets has also changed dramatically. As of 10 October 2018 these had risen to 1,473, or over fifteen times the number proposed by Zorc.¹⁸ The great majority of these have no connection with trade, including nouns such as *balinu 'beach morning glory', *butúl 'stone of fruit', *dúdun 'grasshopper', or *ipus 'tail', verbs such as *atúban 'to face, confront', *bagut 'to pull out, as hair', *balud 'to bind, tie up', *biklaj 'to spread out, unfurl', *bunun 'to distribute, share', or *denđen 'to boil vegetables', adjectives such *dakél 'big', or *hadawiq 'far, distant', or grammatical formatives such as *ka- 'mate, partner', all of which (along with many others) are found in

14 Lawrence A. Reid, 'The early switch hypothesis: Linguistic evidence for contact between Negritos and Austronesians', *Man and Culture in Oceania* 3 (1987): 41–59.

15 R. David Zorc, 'The genetic relationships of Philippine languages', in *FOCAL II: Papers from the Fourth International Conference on Austronesian Linguistics*, ed. Paul Geraghty, Lois Carrington, and S.A. Wurm (Canberra: Dept. of Linguistics, ANU, 1986), pp. 147–73.

16 Malcolm Ross, 'The Batanic languages in relation to the early history of the Malayo-Polynesian subgroup of Austronesian', *Journal of Austronesian Studies* 1, 2 (2005): 12–13.

17 Reid, 'The early switch hypothesis', p. 96.

18 Blust and Trussel, ongoing.

Yami of southeast Taiwan, as well as in languages of the Philippine archipelago. No plausible borrowing hypothesis can account for these or hundreds of other cognate sets that are widely distributed in Philippine languages and unknown elsewhere. Some of these may prove to have external cognates, but the current number of forms makes it unlikely that the Philippine-only lexical database will contract substantially.

Apart from this empirical support for Proto-Philippines, the borrowing hypothesis for Philippine-only lexical distributions also raises difficult theoretical questions, the most pressing of which is why would the rampant multidirectional borrowing that Reid claims to have taken place in the Philippines not have occurred to the same extent in other parts of the Austronesian world?

Chapter 7, 'Macrophyletic trees of East Asian languages re-examined' by Weera Ostapirat, re-examines the linguistic relationship among the five major language stocks of East Asia: Sino-Tibetan, Austronesian, Austroasiatic, Kra-Dai, and Miao-Yao. At 15 pages this chapter is short, but surprisingly informative. After a brief survey of earlier views on linguistic macrofamilies in East and Southeast Asia, Ostapirat presents a set of 24 meanings, drawn from earlier publications that have attempted to establish the most stable parts of basic vocabulary, and he lays out their translation equivalents in Tibeto-Burman, Old Chinese, Austronesian and Kra-Dai. The results show a clear division into two families or super-families: Sino-Tibetan and Austro-Tai, effectively eliminating the claim that either Austronesian or Kra-Dai are genetically related to Sino-Tibetan. His conclusion flies in the face of mainland Chinese scholars, who point to an extensive vocabulary shared by Thai and Chinese, but Ostapirat deftly deals a death blow to this position by showing that the most basic forms in Kra-Dai favour an Austronesian connection, and that lexical connections with Chinese almost certainly reflect a history of heavy borrowing. In the last section of his chapter he suggests that Miao-Yao (Hmong-Mien) may be distantly related to Austroasiatic, a suggestion that, if confirmed, would add further support to an earlier proposal of Jerry Norman and Mei Tsu-Lin.¹⁹

Chapter 8, 'The family tree model and "dead dialects": Eastern Middle Iranian languages', by Yutaka Yoshida, translated from the Japanese by Ritsuko Kikusawa, is an account of problems in the comparative study of the Iranian languages which draws heavily on philological materials. Apart from providing an excellent introduction to Iranian linguistics for the non-specialist, it is primarily concerned with the problem of cross-cutting innovations addressed in more formal terms by Kalyan and François in chapter 5. Perhaps its most striking feature is what Yoshida calls the 'Grimm's Law' of the East Iranian languages, a change whereby voiced stops were spirantised before a vowel and voiceless fricatives were voiced before /t/. Since this innovation is shared by all East Iranian languages, parsimony would dictate that it preceded the break-up of Proto-East Iranian. However, enter philological data and this assumption is contraindicated, since the language of the *Avesta*, the sixth century BC sacred texts of Zoroastrianism, shows no such change, and this language

19 Jerry Norman and Mei Tsu-lin, 'The Austroasiatics in ancient South China: Some lexical evidence', *Monumenta Serica* 32 (1976): 274–301.

is generally assigned, on the basis of place names, to the East Iranian group, making it appear that the East Iranian ‘Grimm’s Law’ is a product of drift.

Chapter 9, ‘What the tree model represents: Language change, time depth, and visual representation’ by Ritsuko Kikusawa defends the traditional position that the family tree model is most useful in macro-comparison, or when applied to more distant relationships, while various diffusion models apply more appropriately in micro-comparison (dialectology). While this view is not original, she is original in applying it to both spoken and signed languages, noting that relationships between the latter cannot be represented by a family tree, as they show a reticulate pattern due to borrowing (for example, American Sign Language is more closely related to French Sign Language than it is to British Sign Language, reversing the relationships seen among the corresponding spoken languages).

Some features of this chapter may raise eyebrows, as the implication that Otto Schrader was the first critic of the classic family tree model of August Schleicher, rather than Johannes Schmidt.²⁰ The Austronesian family tree presented in her fig. 9.5 is not the consensus view, but is pieced together from various sources in a very idiosyncratic manner, and the statement that ‘The existence of “similar” languages covering a vast area reaching from the western edge of the Indian Ocean to the eastern Pacific was recognised by European travelers from the beginning of the seventeenth century’ (n.1) is inaccurate. The relationship of Malagasy to Malay was recognised in 1603, the connection of these languages to those in western Polynesia followed in 1615, and remained unchanged when Hadrian Reland recognised a ‘common language’ reaching from Madagascar to western Polynesia nearly a century later.²¹ It was only after the three voyages of James Cook from 1768 to 1779 that the Spanish scholar Lorenzo Hervas y Panduro recognised a ‘common language’ extending from Madagascar to eastern Polynesia.²²

In conclusion, I found this volume useful, although it may give the misleading impression that it is addressing a problem that historical linguists have been unaware of until now. The book is well-edited, with relatively few typos or other mechanical errors. One that stands out is Minaka’s translation (p. 10) of the German (‘*Wir müssen gruppieren*’) as ‘we cannot make groups’, which precisely reverses the intended meaning.

20 Otto Schrader, *Zur Geschichte und Methode der linguistisch-historischen Forschung* (Eschborn: H. Costenoble, 1907); Johannes Schmidt, *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen* (Weimar: Hermann Böhlau, 1872).

21 Hadrian Reland, *Dissertationum miscellaneorum*, 3 vols. (Trajecti ad Rhenum, 1708).

22 Lorenzo Hervas y Panduro, *Catalogo delle lingue*, vol. 17 of *Idea dell’Universo*, 21 vols. (Cesena, 1784); Robert Blust, *The Austronesian languages*, rev. ed., Asia-Pacific Linguistics Open Access Monographs (Canberra: College of Asia and the Pacific, ANU, 2013 [2009]).