Quantifying uncertainty in the phylogenetics of Australian numeral systems

Kevin Zhou1 and Claire Bowern2

1Department of Biomedical Engineering, Yale University, 55 Prospect St, New Haven, CT 06511, USA
2Department of Linguistics, Yale University, 370 Temple St, New Haven, CT 06511, USA

Researchers have long been interested in the evolution of culture and the ways in which change in cultural systems can be reconstructed and tracked. Debates about the ways in which such systems change have focused on the relative frequency of horizontal diffusion versus vertical transmission. That is, are cultural traits primarily inherited, or do they spread across geographical areas [1–4]?

Within the realm of language and culture, these questions are increasingly investigated with Bayesian phylogenetic methods. Recent research has investigated questions in kinship [5,6], basic vocabulary [7–11] and grammatical patterns [12,13]. Such work uses Bayesian inference to design and test evolutionary scenarios. But results are often presented without full and explicit quantification of the uncertainties associated with models, reconstructions and trees used to formulate descent hypotheses. For example, some papers [13] use tree samples for their phylogenetic trait analysis, yet present only a single model for transition probabilities, without quantifying the support for that model or discussing alternatives. Such work, important though it is, could be improved by more explicit quantification of reconstruction and transition probabilities.

We address these issues by developing a series of methods for quantifying and visualizing results of ancestral trait reconstruction. We test them using data from numeral system reconstruction in Australian languages of the Pama–Nyungan family [14]. Numerical systems across the world show great variation [15–17], but within Australia, the systems tend to be low limit (with an upper bound of less than 20, and frequently less than 5). However, while there is variation in both the composition and the extent of the systems, that variation is highly constrained. Low-limit numeral systems are not unique to Australian languages, but are characteristic of hunter–gatherer societies throughout the
world, which presumably have a lesser functional need for a more general system [15]. Variation in Australian numeral systems stems not only from the systems' upper extents but also from whether larger numbers are composed of smaller numbers (e.g. \(3 = 2 + 1\), \(4 = 2 + 2\), etc.; compare English 'twenty-five') or show unanalysable, opaque etymology (such as English 'three' and 'four') [15,16]. While previous work [16,17] recognizes some variation in the extent of low-limit numeral systems, there has not been any investigation into how low-limit systems change over time, particularly with respect to their upper limits. Do such systems gradually increment, for example? And can such systems lose numerals, and revise their upper limits downwards? Finally, Australian languages are also claimed to show extensive horizontal transfer of features [18], but Bayesian phylogenetic trait analyses allow us to investigate the level of phylogenetic signal in a dataset. All these factors make Australian numeral systems an ideal test case for phylogenetic investigation.

2. Methods

(a) Numerical systems

We investigate two questions about the evolution of Australia's numeral systems. First, we explore the phylogenetics of numeral system limits across the Pama-Nyungan family (the most populous family of Australian languages, covering almost 90% of the land mass and containing approx. 70% of Australian languages). Bowern & Zentz [14] show that upper limits vary from 3 to 20, with the median at 4. We investigate whether limits are structured phylogenetically. Within this domain, we test four hypotheses about numeral evolution. One hypothesis is that languages add numerals incrementally over time. If this is so, we would expect the root (Proto-Pama-Nyungan) to reconstruct to 2, and for numeral extent to increase serially approaching the tips. A second hypothesis is that upper limits are random, or the products of local system enhancement. A third hypothesis is that languages with low-limit numeral systems may lose numeral terms over time (an alternative hypothesis to hypothesis 1 that limits strictly increase). A fourth hypothesis is that languages might borrow numerals from one another, thus leading to horizontal transfer of features of numeral systems as well as of the words used in those systems.

Secondly, we investigate whether numerical creation and replacement evolves phylogenetically. Languages vary in whether their numerals are compositional or opaque. Compositional numerals are formed from two other numbers (e.g. English 'twelve-one' = 'twenty-one') and 3 is buula-nguunbu (two' + 'one') and 4 is buula-buula ('two' + 'two'). There is substantial variation across the languages, however, as to whether both 3 and 4 are compositional or opaque, and if only one is compositional, which numeral shows the compositional form. We investigate this further by comparing evolutionary models where the form of 4 is dependent on the form of 3, versus models where the two numerals are assumed to change independently. We further hypothesize that opaque numerals will tend to be renewed or replaced by compositional forms.

(b) Bayesian statistics

Bayesian phylogenetic methods are relatively new to the study of linguistic prehistory, serving as a complement to the comparative method [19,20]. They were adapted from methods developed for evolutionary biology upon the realization that many of the properties of evolution have direct parallels in language change [7,21]. These statistical methods, particularly Markov chain Monte Carlo (MCMC [22]), have been successful in verifying conclusions drawn from more traditional linguistic methods as well as offering support for deeper and less obvious relationships (e.g. [10,11]). As is well known, MCMC traverses not only the parameter space but also the model space. In this way, models that poorly explain the data are visited less often. In our case, the parameters are transition rates between the different states (i.e. numeral extent or opacity) throughout the language tree, as well as the state reconstructions at various nodes within the tree, which are represented as probability mass functions, allocating different probabilities to each state in proportion to the degree of their support. We use an extension of MCMC known as Reversible Jump MCMC (RJMCMC; [23], see [24] for its introduction to (and development in) evolutionary biology), which allows the number of parameters to change. This technique is particularly suited for evolutionary models with a large number of possible states. RJMCMC traverses not only the parameter space but also the model space. In this way, models that poorly explain the data are visited less often.

In our case, the parameters are transition rates between the different states. Given a set of \(k\) independent traits, there are \(k(k - 1)\) transition rates. However, calculating all of the transition rates is not only computationally demanding but also unnecessary because it leads to overparametrization. To reduce the number of transition rates, we can either delete rates or restrict them to the values of other rates, but there is no systematic method of determining these deletions and groupings without external knowledge. RJMCMC overcomes this issue by allowing the Markov chain to 'jump' from one dimension to another by deleting or grouping parameters together in addition to varying them. As with conventional MCMC, such jumps are more likely to occur if the candidate parameter vector better fits the data. Hence, we can draw conclusions about the different transition rates based on, for example, how often they are deleted.

To model numerical system extent, we use BayesTraits [25] to infer ancestral states at different nodes. To investigate correlated evolution for numerical opacity, we examine two classes of models, one that assumes independent evolution and one that assumes dependent evolution, which we then compare via the Bayes factor (BF) [26]. Fitting the data to both models and comparing their posterior likelihoods via the BF indicates whether coevolution is favoured. We adopt the RJMCMC approach, but we modify the coevolution model of opacity of '3' and '4' by introducing two new paired states (electronic supplementary material figures S1 and S4). These are required because the opacity of '4' is ternary—opaque, compositional, or not attested. We retain the opacity of '3' as a binary term because no languages have a term for '4' but not for '3'.

(c) Representation of RJMCMC reconstructions

Given that each RJMCMC run fits the data to multiple models, there is potential for the state reconstructions to vary more significantly than those deduced by conventional MCMC methods, as we shall see. For example, there might be two significantly different sets of models that both explain the data well. As such, representing the uncertainty in the reconstructions using a simple mean and confidence interval in place of a histogram of the reconstructions would obscure the true nature of the reconstructions. Here, we propose a new method that uses the Shannon entropy of the reconstructions as a sorting criterion. Briefly, the entropy of a discrete trait X with a probability mass reconstruction \(p(x)\) is given by \(H(X) = - \sum_x p(x) \log p(x)\). If \(p(x) = 0\) for some \(x\), then \(p(x) \log p(x)\) is defined to be 0. In most
applications, \(k = 2\) and the units are in bits; however, for convenience we assume \(k\) to be the number of possible values \(X\) can take so that the entropy is bounded between 0 and 1. Entropy has many interpretations depending on the application, but here it is useful to think of it as how certain the reconstruction is. For example, an entropy of 0 indicates that one particular value of \(X\) has a probability of 1, meaning that the reconstruction is completely certain. On the other hand, a maximum entropy of 1 indicates that the reconstruction is uniform across all the states—that is, it is completely uncertain.

By sorting reconstructions based on their entropies, we achieve a better sense of the diversity of the reconstructions returned by RJMCMC. Furthermore, as we will see, this representation is more compact than using histograms and provides more information. Generating one-dimensional histograms is an inherently marginalizing procedure—in other words, it discards any joint relationships. Our procedure displays the joint reconstruction and simultaneously visualizes all the reconstructions given by the Markov chain. Also of interest is the distribution of the entropies themselves, the shape of which conveys the overall degree of certainty the procedure was able to confer. Hence, we weighted each sorted entropy by how frequently each reconstruction of that entropy appeared. Doing so accentuates the reconstructions and the entropies that appear most often. Estimates of the entropy appeared. Doing so accentuates the reconstructions and the entropies that appear most often. Estimates of the entropy density were given by spline interpolation of histograms.

(d) Analysis of rates

To investigate the transition rates that figure importantly in the analyses, we rank the rates. Ranking rates by their speed across the MCMC chain allows us to quantify which rates are relatively fast: for each state in the RJMCMC chain, we ranked the rates from fastest to slowest and tabulated the proportion of the chain each rate was fastest. We also plot (for comparison) the proportion of runs for which each rate is deleted. Doing so exploits the joint Bayesian relationship among all the parameters, while merely comparing the mean rate, as some authors have previously done [24], ignores the joint relationship.

3. Results

(a) Extent

Figure 1 displays the entropy-sorted reconstructions at various ancestral nodes throughout the tree. The reconstructions vary in certainty at any given node and across all nodes. The closer to zero the reconstructions cluster on the \(x\)-axis, the lower the entropy and the greater the confidence in the reconstruction. For example, the Mayi and Ngayarta subgroup node reconstructions have entropies approaching 0, where an extent of four and three (respectively) are confidently reconstructed. On the other hand, the Central New South Wales node is not as confidently reconstructed, as illustrated by the relatively high entropy levels, even though reconstructions of an extent of 4 predominate.

Marmgu provides an illustration of conflicting models. Lower entropy reconstructions favour a limit of three, but the peak density for reconstructions indicates greater overall support for limit-five, although with less certainty. Likewise, Paman ancestral state reconstructions suggest two solutions; either limit-three or limit-five. Since reconstructions of limit-three are allocated a higher posterior probability across the entropy range, they are to be preferred. The root (Proto-Pama-Nyungan) reconstruction appears to be limit-four. Certainty of the reconstructions is affected by the degree of posterior support for the particular node in the tree. As Pagel et al. [25] put it, ‘the probability that the node exists puts an upper limit on how certain one can ever be about the ancestral state of that node’. Most of the nodes under study here have probabilities in the sample at or close to 1; however, it should be borne in mind that trait reconstruction confidence can never exceed the probability of the node it is assigned to.

Low-limit numeral systems do not only increase their limits over time, and they do not simply increase limits serially. Languages can also lose numerals, supporting hypothesis 3, and contrary to strong versions of hypotheses 1 and 2. In the Central subgroup of Pama-Nyungan, for example, the most likely reconstruction is limit-four, but Karnic, one of the daughter subgroups, has substantial support for limit-three. Several subgroups, including Paman and Western Pama-Nyungan, show support for a reconstruction limit of three, in comparison to the root, with limit-four.

Transition rates provide further information about which changes in extent are particularly common. A transition rate from numeral extent \(x\) to \(y\) is denoted by \(q_{xy}\). Thus, \(q_{34}\) is the transition rate from state 3 to state 4 (or, in real-world terms, the transition from a language having three numerals to having four). Rates that are frequently deleted have less importance in the model. We rank rates by speed and plot two pieces of information: the proportion of runs in which the rate is the fastest, and the proportion in which the rate is deleted (figure 2). Three rates stand out as particularly fast: \(q_{34}, q_{35}\) and \(q_{43}\). This implies that in our data, limited elaboration of the numeral system is common (from extent 3 to extent 4 or 5), as is some simplification (from extent 5 to 4). These results also imply that strict stepwise elaboration of low-limit systems is not necessary, given the rate of extent elaboration from 3 to 5.

The next fastest rate is \(q_{49}\). This perhaps implies that there is a threshold for numerical complexity (that is, if elaboration beyond 4 or 5 is found, significant rather than marginal elaboration is found). This accords with research on acquisition of numerals [27,28], which shows that children encode the numerals involved in subitizing (\(1–3\)) earlier. They map small numerosities to specific lexical items and only later generalize to higher numerals [27]. Rates \(q_{37}, q_{47}\) and \(q_{57}\) are overall slow and are most frequently deleted. Stepwise rates (e.g. \(q_{34}, q_{45}\)) are, on the whole, faster than rates where limits increase or decrease non-incrementally (e.g. \(q_{35}, q_{24}\)), implying that while hypothesis 1 is false, stepwise incrementation is still more common than non-stepwise elaboration. These data are consistent with a situation in which numeral extent rates fluctuate stochastically, but where variation in numeral extent is mostly (though not entirely) concentrated around accretion and deletion of numerals between 3 and 5. Hypothesis 3, that languages can lose numerals, even in low-limit systems, is thus well-supported in our dataset.

The final hypothesis concerns the borrowing of numeral systems. The data suggest elaboration of numeral systems in the languages of the South subgroup (Kulin, Waka-Kabi, Central NSW and Lower Murray). While there is no direct evidence from the forms of the numerals that borrowing has taken place, it is striking that the only subgroup outside the Southern clade with numeral extent above 5 is Thura-Yura, a subgroup that is directly adjacent to the Lower Murray languages. Wati, the subgroup with the next highest extent reconstruction, borders Thura-Yura to the west. This implies that horizontal transmission may well play a role in the elaboration of numeral systems, supporting hypothesis 4.
We next consider the internal structure of numeral systems. As described above, languages vary in the extent to which higher numerals are built from smaller numerals; that is, whether they are compositional, or opaque. One obvious question is whether the compositionality of higher numerals depends on that of lower numerals; for example, if the word for ‘4’ is based on 2 + 2, does that imply that the word for ‘3’ is also compositional (based on 2 + 1)? Compositionality changes over time, through a process known in linguistics as grammaticalization [29]. As shown in electronic supplementary material, figure S1, all possible combinations of opaque and compositional numerals exist in our dataset, though at unequal frequencies.

To test for correlated evolution of compositionality, we ran two models. In the first, numerals were assumed to evolve independently. In the second, they were assumed to evolve dependently. The BF was then compared to evaluate the extent of support for one model compared with the other. A BF range of 7.8–10.4, verified from three independent Markov chains, indicates that the dependent model is superior to the independent model, offering strong support for correlated evolution between the opacity of ‘3’ and ‘4’.

Electronic supplementary material, figures S9 and S10 give...
the entropy-sorted reconstructions at ancestral nodes using both models.

If we compare the reconstructions produced by the dependent and independent models, we note that some reconstructions are similar (e.g. Paman, Kalkatungic) while others are different (e.g. Root, Kulin and Mayi). The South node is interesting in that although state 1 (compositional 3 and 4) has the greatest a posteriori probability under both models, the dependent model sometimes gives state 5 (i.e. compositional 3 and absent 4) as a strong alternative. Furthermore, the entropy ranges differ; in general, it appears that the dependent model produces reconstructions with a broader range of entropies, often extending closer to 0. It also tends to produce results without a clear reconstruction across entropies. Furthermore, if we compare the reconstructions with those considering extent data (without regard for numeral etymology, from figure 1 above), we see that there is overall consistency between the two sets of reconstructions, and subgroups reconstructed with extent 3 tend to be reconstructed with absent 4. Within Western Pama-Nyungan, for example, both Warluwaric and Yolngu reconstruct with high probability to extent-3 numeral systems. For Warluwaric, compositional 3/absent 4 is the most likely node value, while for Yolngu opaque 3/absent 4 is most likely.

To get a sense of the mechanism of correlated evolution, we can compare the various transition rates of the dependent model. Figure 3 summarizes the reconstructed transition rates in the dependent model, showing particular dependencies. Arrows are weighted by frequency of the rate occurring in the sample. Therefore, the thickest arrows (e.g. between 0,0 and 0,1 where opaque 4 becomes compositional) show the transitions that are most relevant for the model.

The data provide evidence for the ways in which languages both lose and gain numerals. The three fastest rates involve the creation of a compositional 4 either through replacing an earlier opaque 4 or incrementing a smaller system. Note, however, that an opaque 4 may also be both gained or lost. This behaviour is in contrast to 3 which, in our data, becomes opaque only if 4 is also opaque. This is presumably in large part because compositional 3 is usually ‘2 + 1’; this leads to a pressure for 4 to be compositional in base 2 as well (i.e. 2 + 2; see further discussion of this by Bowern and Zentz [14]). These findings make sense in a model of language change where new elements in a numeral system are created and renewed analogically to other numerals. An alternative would be where speakers of languages recruit opaque terms from other domains [15]. This is the usual source of words for ‘5’, which frequently originate in words for ‘hand’ in Pama-Nyungan languages (infrequently, words for 5 based on ‘3 + 2’ are found). This is not the dominant mode of evolution of terms for 4 in Pama-Nyungan languages, however. These data also imply a cycle of grammaticalization where numerals (particularly 4) tend to be innovated compositionally, grammaticalized or fossilized so that their origin becomes opaque to their speakers, then lost or replaced by a new compositional form.

Compositional 3 (e.g. 2 + 1) and opaque 4 seems to be an unstable state, with fast transition rates leading to loss of 4, replacement of opaque 4 with compositional 4, or replacement of compositional 3 with opaque 3. Finally, 4 may be gained or
lost, whether 3 is opaque or compositional. This implies that the upper limit of a numeral system is, at least to some extent, independent of numeral etymology, as one would expect if these are true numerals. If the numeral words in these languages are \textit{ad hoc} compounds rather than true numerals, we would expect there to be a clearer dependency between the presence of a compositional 3 and a compositional 4.

4. Conclusion

In conclusion, entropy sorting provides a convenient way to visualize the relative likelihoods of reconstructions of multistate data at different nodes in the phylogeny. Both entropy plots and rate comparisons allow a way to evaluate the outcomes of Bayesian phylogenetic analysis. We use this approach to investigate the phylogenetics of small numeral systems in Australia. We reconstruct the most likely extent of the system at the root and use that information to explore the ways in which numeral systems evolve. We show that the mode of evolution among low-limit numeral systems in Australia is variation between limits of 3, 4 and 5, and that languages may both lose and gain numerals across time. We provide evidence for a rapid elaboration of numerals above 5, and that there is some evidence that higher numeral extents have spread areaally. Below extent 5, however, languages fluctuate in their upper limits across time. These findings argue against a model of numeral evolution which begins with gradual incrementation or elaboration. We investigate the etymologies of numerals and show that numeral 4 can be lost or gained independently of the structure of 3, numeral 4 is usually innovated compositionally.

Data accessibility. See electronic supplementary materials for coded datasets and references to original sources.

Authors’ contributions. C.B. and K.Z. developed the research; K.Z. wrote the code for the analysis; C.B. sourced data; K.Z. and C.B. analysed the data; C.B. and K.Z. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. This work was funded by NSF grants BCS-0844550 and BCS-1423711.

Acknowledgements. Thanks to members of the Pama-Nyungan/Historical Language Lab at Yale, and to Mark Pagel.