**Solving Galton's problem: practical solutions for analysing language diversity and evolution**

Lindell Bromham
Macroevolution & Macroecology group,
Division of Ecology & Evolution,
Research School of Biology,
Australian National University,
Canberra ACT 0200
lindell.bromham@anu.edu.au
tempoandmode.com

**Abstract**

Comparisons between languages can illuminate the processes of language change, by revealing meaningful associations between language features, or the influence of external factors on the patterns and rates of language change. But comparisons between languages raise statistical challenges, because close relatives will tend to be more similar to each other in many respects than they are to more distantly related languages, and languages from the same regions will be subject to many of the same influences. Therefore observations made on different languages will usually fail to meet the requirement of statistical independence inherent in all standard statistical tests. This fundamental challenge of cross-cultural analysis, known as Galton's problem, is no cause for despair, because there are a range of solutions that fit comfortably within the scope of most projects, using only data that are easily available for all languages. This paper discusses a range of practical solutions, including phylogenetic analysis, sister pair comparisons, and spatially structured models, that can be applied to analyses of language variation and change.

**1. What is Galton's problem?**

Francis Galton's contributions to statistics grew from his fervent desire to understand human diversity (Bulmer 2003). Galton was an obsessive measurer, committed to developing better ways to extract meaning from data, and constantly seeking clever ways of doing just about everything (take, for example, his description in *Nature* of a scientifically principled way to slice a cake (Galton 1906)). It was Galton who developed the modern concept of statistical correlation. His goal was understanding patterns of heritability in humans, but, for the sake of practicality, Galton first used sweet-peas as an easily obtained substitute that was, helpfully, uniparental (Stanton 2001). He found that plotting the weight of the parent seed against the weight of their offspring seed gave a straight line. A horizontal line would suggest no particular relationship between parent and offspring traits, but a slope less than one showed a correlation between parent and offspring, because the weight of the parent seed provided some information on the likely value of the weight of the offspring seed. In other words, these results provided evidence that seed size was inherited from one generation to the next.

Francis Galton adopted the term "co-relation" to express the way that body parts vary together – for example, people with long legs also tend to have long arms, so the correlation coefficient describing the relationship between the measurements will be between one (complete association between values) and zero (no association in values). "It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to the common causes. If they were wholly due to common causes, the co-relation would be perfect… If they were in no respect due to common causes, the co-relation would be nil. Between these two extremes are an endless number of intermediate cases, and it will be shown how the closeness of co-relation in any particular case admits of being expressed by a simple number" (Galton 1889b). Galton developed correlation analysis because he wanted to understand biometry (the relationship between traits within an individual) and inheritance (the relationship between traits among relatives), and to do this he needed to describe the dependence of values on each other (Denis 2001; Pearson 1920).

The familiar mantra "correlation doesn't prove causation" may be true, but it is somewhat specious, because uncovering causal connections is precisely what correlation analysis is designed for – identifying cases where the association between two or more variables is greater than would be expected by chance, and therefore indicative of a link between the two (Shipley 2002). However, correlation analysis does not reveal the nature of the causal connection between variables. Galton pointed out that correlation between variables could be caused by inheritance from a common ancestor, a shared environment or common history – an effect now known as 'Galton's problem'

following a response that he gave to a presentation at the Royal Anthropological Institute in London in 1888 (Naroll 1965).

At that meeting, anthropologist Edward Tylor presented quantitative analyses of an impressive database of hundreds of cultures, which he used to identify associations between marriage practices and other features of social organisation, for example residency patterns and avoidance customs (Tylor 1889). But Galton pointed out that shared inheritance provided an alternative explanation for the correlation between marriage practise and other features of social organisation: if multiple cultures in the sample have the same marriage practice because they all inherited it from the same ancestor, then they may have also co-inherited other cultural traits, such as residency patterns. Traits may be associated with marriage practice not because there is any meaningful connection between them, but because related cultures inherit many different features from the same shared ancestor (Galton 1889a). In Galton's words "It was extremely desirable for the sake of those who may wish to study the evidence for Dr Tylor's conclusions, that full information should be given as to the degree in which the customs of the tribes and races which are compared together are independent. It might be, that some of the tribes had derived them from a common source, so that they were duplicate copies of the same original. Certainly, in such an investigation as this, each of the observations ought, in the language of statisticians, to be carefully "weighted". It would give a useful idea of the distribution of the several customs and of their relative prevalence in the world, if a map were so marked by shading and colour as to present a picture of their geographical ranges" (Galton 1889a).

**2. Who has Galton's problem?**

All standard statistical tests assume independence of datapoints. For a set of observations, statistical tests such as correlation analyses and t-tests ask whether if you know the value of one variable for a given observation, you can predict the value of another. For example, our observations might be made on a set of languages, so each language is one datapoint, and for each language we might ask if knowing the population size allows us to predict language complexity. For a fair statistical test of the relationship between variables, it is important that the observations you make are independent of each other – in other words, statistical independence is a requirement of datapoints, so that we can properly test the dependence between variables. Galton was concerned that Tylor's observations on the correlates of marriage practice would be non-independent if knowing the marriage practice in one culture would allow you to make an informed guess at the marriage practise in a neighbouring or related culture. For observations on different cultures to be statistically independent, related or neighbouring cultures should be no more likely to be similar to each other than a randomly chosen, unrelated cultures. The same problem applies

to any analysis of entities related by descent, and is familiar to evolutionary biologists, who refer to the problem as "phylogenetic non-independence".

The reason that this is important is that we want analyses of cross-cultural observational data to be like a well-designed experiment that asks the question: if there is a change in one cultural trait, will it results in change in other cultural traits? (Bromham 2016). For this test to work, we need to know that the traits we are interested in are free to vary, so that each observed incidence of co-occurrence is evidence of a change in one variable causing a change in the other. But if the cultural factor we are interested in is heritable (usually similar to its immediate ancestor), then it will also tend to be similar amongst relatives. In this case, the value of the variable amongst datapoints is not independent, because knowing the value of one gives you information on the likely value of its close relatives. When we observe that many related cultures have the same combination of traits, we may be recording the outcome of a single evolutionary event, not a repeated trial. In Galton's terms, making observations on related cultures is like taking duplicate copies of the same original. Enumerating the cases where certain features co-occur (as Tylor did for marriage practices) may result in essentially counting the same observation over and over again. This "pseudoreplication" inflates the appearance of a significant association between traits, even if they are not functionally connected. If relatives tend to have similar values for inherited cultural traits, then it will often be the case that they will share similar values for many different variables: they will tend to have similar marriage practices and similar residency patterns, even if there is no connection between the two.

As a simple illustration of the problem of non-independence of observations, consider the association between chocolate consumption and Nobel prizes. A significant correlation has been reported between per capita chocolate consumption per country and the number of Nobel prizes its citizens have been awarded (Messerli 2012). The correlation tells us that high values of chocolate consumption tend to be found in countries that also have high values for Nobel prize success - but what is causing this association? It could be that the flavanols in cocoa result in cognitive enhancement (Sokolov et al. 2013) and thus accelerate innovation and creativity. Or it could be that the countries with high chocolate consumption incidentally all share some other feature that makes Nobel prizes more likely. Most of the highest chocolate-consuming countries are in northern Europe. These countries also share a higher-than-average proportion of all Nobel prizes awarded: Northern Europe has ten times more Nobel prizes per capita than the world average. For example, Sweden and Switzerland have gained 58 Nobel prizes between them, while China and India have had only 20 Nobel prizes between them, despite having 150 times more people. Maybe this is because Northern Europeans eat more chocolate, stimulating their national intellectual prowess. But it might also be because they spend more on research per capita

(Folyovich et al. 2019): Switzerland and Sweden ranked second and fifth highest research expenditure per capita, China and India ranked 37th and 73rd. The correlation between chocolate consumption and Nobel prizes is true – these variables are associated with each other when compared across countries – but the causal link may be due to similarity of related cultures rather than a direct connection between chocolate and Nobel-worthy work. In fact, nearly any characteristic of northern European countries might also be found to correlate with both chocolate consumption and Nobel prizes. For example, the number of IKEA stores per capita is correlated with Nobel prize success (Maurage et al. 2013).

What does this mean for linguistics? Closely related languages can vary in lexicon and grammar, but, on the whole, we expect a language to be more similar to its close relatives than it is to a language chosen at random by a blindfolded throw of a dart at a world map. These shared features between relatives are the fundamental basis of the comparative method in historical linguistics, as well as language phylogenetics (Atkinson and Gray 2005; Rankin 2003). But they also create patterns of covariation in the data. In the process of language evolution and diversification, many features will be inherited together, not because those features are fundamentally connected, but because features that happened to co-occur in an ancestral language are now also found in many of their descendants. Any analysis that uses observations from countries or languages as datapoints needs to account for the similarity between relatives and neighbours in order to identify meaningful patterns of association between language features.

Take, for example, an important study examining the relationship between speaker population size and language complexity for a sample of over two thousand languages (Lupyan and Dale 2010). By regressing population size against several measures of complexity, such as number of categories per verb and number of nominal cases, they conclude that large populations have simpler language structures, and they hypothesise this is due to the simplifying effect of language contact and larger numbers of L2 speakers. But, just as related languages often have similar grammatical structures, they also have a tendency to be more similar in population size that you would expect from a random draw from all languages. This does not mean that we expect close relatives to always be similar in population size. But if you lined all the world's languages up in order of their speaker population sizes, you would not expect all language families to be evenly distributed throughout that distribution. Although there is a wide range of population sizes in both Pama–Nyungan (Australian) and Indo-European languages, the distributions of population sizes from each family will barely overlap. This sets up the conditions for any consistent difference between these families to incidentally associate with differences population size. For example, 48 languages in WALS are recorded as having no fricatives, and the majority of these languages are in Australia, where speaker population sizes are relatively low (Maddieson 2013a). We should not

be surprised, therefore, if lack of fricatives is associated with population size, but such a relationship would not necessarily tell us anything useful about the evolution of consonant inventories.

We can imagine many other families would similarly cluster along the size continuum, and therefore carry a possibility that different population sizes will associated with different language features. Because Indo-European languages tend to have larger population sizes than American or Australian languages, anything that differs consistently between European languages and American or Australian languages could also associate with population size. If each Indo-European language is counted as an independent instance of the association between complexity and population size, and each Australian or American language is included as separate datapoints, then there is a risk that inherited aspects of language complexity will appear to be associated with consistent differences in speaker population size (Figure 1). The same problem will also occur at sub-family comparisons, if language groups within a family tend to be more similar in both population size and features of language complexity. For example, within the Austronesian language family, we might expect languages within different sub-families to be more similar to each other both in terms of population size and features of language complexity than they are to more distantly related members of other subfamilies. We would expect Austronesian languages in South East Asia to differ in many ways from Austronesian languages from Polynesia, including both languages features, population characteristics and environmental conditions. This is why adding language family as a factor in the analysis does not remove the problem of non-independence due to shared history (Koplenig 2019), because languages within families will still show patterns of covariation due to their relationships to each other. It is important to emphasize that non-independence due to inheritance does not necessarily mean the conclusions are false, but it does mean that we can't use the observed correlation to support that a functional connection between language features and population size until we can discount association due to inheritance.

It is not just history and relatedness that can generate patterns of non-independence in comparative analysis of language diversity. Languages that cluster together in space can also show patterns of non-independence due to their shared environments. For example, a correlation between parasite load and language diversity has been interpreted as supporting the hypothesis that infectious disease risk has played a major role in shaping human cultural diversity, on the assumption that populations living under high parasite load will benefit from limiting contact with other populations (Fincher and Thornhill 2008). Parasites show a strong latitudinal diversity gradient, as do species that act as sources and reservoirs for human disease (Dunn et al. 2010; Nunn et al. 2005; Poulin 2014). This means that anything else that shows a latitudinal gradient,

such as GDP, temperature and language diversity, is also likely to correlate with parasite load (Bromham et al. 2020; Collard and Foley 2002; Guernier et al. 2004; Kummu and Varis 2011; Mace and Pagel 1995). In fact, number of birds and mammal species provides a stronger predictor of language diversity than parasite load does (Bromham et al. 2018b). This correlation between languages and biodiversity is unlikely to be due to a direct association, but arises because both covary with environmental parameters (Hua et al. 2019).

Non-independence due to shared inheritance is a problem for any analysis of evolved entities, including species, languages and cultures. Many approaches to solving this problem have been developed. Some methods explicitly model the likely degree of covariation due to descent (see section 3.1 below). Other methods rely on selecting datapoints that represent statistically independent observations (Figure 2, section 3.2). A common way to do this is to identify independent contrasts between pairs of languages (3.3). If two languages have descended from a common ancestor, then each began with the same starting value for all features – grammar, lexicon, population size etc – so any difference between them has evolved since then. Therefore we can compare the difference between them and look for associations between language features – does the language with the larger population size also have a simpler grammar? One such comparison does not make a robust test of a hypothesis. But if we have many different pairs of relatives, and for each pair we can compare the values of different variables, then we can start to look for generalities – is it usually the case that the larger language in a pair of relatives has simpler grammar? Does this occur more often than we would expect from a chance association? Such comparisons will be statistically independent tests of the hypothesis as long as every contrast has a unique start point (ancestor) and end points (descendants) that are not included in any other comparisons. A practical test of phylogenetic independence of comparisons is to be sure that a line connecting the two members of each contrast on a phylogeny would not cross the line connecting any other contrast – if it does then the same shared history is counted more than once (Figure 2).

If you knew the value of variables at internal nodes of the phylogeny then you could also make comparisons between those (Felsenstein 1985; Harvey and Pagel 1991). Some methods of phylogenetic independent contrasts use a model of change to infer the value at these internal nodes to allow more contrasts to be made, but all such methods rely on the adequacy of the model of change (see below). In addition, the amount of time that has elapsed since the common ancestor and the speed of change will influence how many differences we expect to observe between relatives. Long-separated languages, even if they are each other's closest relatives, will probably be more different from each other than two languages that have recently diverged. A phylogenetically-informed analysis tells us when we should be surprised that two features co-

occur, by comparing the frequency and pattern of co-occurrence given what we know about the potential for shared inheritance, and their opportunity for independent change. If language features co-occur far more often than we would expect on the basis of their shared history, we will need to seek a special explanation, not simply co-inheritance from a common ancestral language.

In considering the functional relationships between language features, it is not simply the co-occurrence of features that is informative, but what happens to one trait when you change the other. The power of a comparative test of association between language features is determined not by the number of observed languages but by the number of inferred state changes. For continuously evolving traits that can be expressed numerically, such as population size, comparisons between nearly any pair of related species provide useful information, because they will probably differ in the trait of interest. But we might have much less explanatory power for categorical variables (such as word order), particularly those that have few possible states (such as having an inflectional future tense or not) or change very rarely (such as whether a language has clicks). For example, if you wish to compare phoneme inventories in languages with and without clicks, the statistical power comes not from how many click and non-click languages you compare, but how many independent origins of clicks you can identify (Daneyko and Bentz 2019; Tishkoff et al. 2007). If many click languages have large phoneme inventories, but they inherited the large phoneme inventories from a common ancestor that had clicks, we can't tell whether clicks and phoneme size are causally related or whether they were both inherited together but not functionally connected (Figure 1). The statistical sample size is determined not by the number of languages but the number of independent origins, or transitions between possible states, because each change in a trait is an experiment on its effect on the other trait. If a language evolves to have clicks will it also increase the phoneme inventory? Or, conversely, is a language with a large phoneme inventory more likely to gain clicks? One state change is one datapoint in the test of this hypothesis.

### 3. Practical solutions to the problem of relatedness

Galton's problem is pervasive in comparative linguistic studies where datapoints are observations made on different languages, cultures or countries. Most researchers are well aware of the problem, and many attempts have been made to ameliorate it. But many of the approaches that have been applied are ineffectual, or partial solutions at best. Selecting a sample of distributed languages or cultures is not a solution to Galton's problem, because we would still expect patterns of relatedness to structure similarity in the sample:

the languages in the sample to be more likely to have similar values to the most closely related languages in the sample (Dow and Eff 2008; Eff 2004). For example, although the Standard Cross-Cultural Sample (Murdock and White 1969) contains non-neighbouring cultures, we still expect cultures within the same group to be more similar to each other than any is to an unrelated culture on the other side of the globe (see Bromham et al. 2018). For the same reason, collecting observations into "bins" or "bands" corresponding to particular regions is not a solution to non-independence, because the observations will still be structured by proximity and relatedness. Nearby languages will tend to be more similar to each other, whether they are within the same "band" or in different bands. For the same reason, adding family as a factor in an analysis does not remove the problem of relatedness, because within any family, variation will still have phylogenetic structure, with close relatives more similar than more distant family members. Adding taxonomic levels as random variables in an analysis is an incomplete solution to the problem of non-independence due to relatedness (Jaeger et al. 2011): significant but misleading correlations between traits may still be driven by patterns of relatedness within taxonomic groups.

Another approach has been to claim that correcting for history and relatedness is not necessary because languages and cultures are so labile and responsive to current environment that we can disregard the possibility of historical inertia influencing observations. This may be true, but it cannot be assumed (Blomberg et al. 2003). An analysis should not be based on the assumption that observations are independent unless it can be proved that relatives are no more similar to each other than expected by chance (which will usually involve some kind of analysis along a phylogeny or taxonomy: see 3.1 and 3.2 below). If relatives are more similar than randomly chosen languages, then this needs to be accounted for in any analysis. In any case, some estimate of relatedness will be required in order to either correct for covariation due to relatedness or to prove that the covariation does not exist for a given dataset (e.g. Roberts et al. 2015). The analysis can be adjusted to the level of information on relatedness that is available for the languages in question, whether that is a phylogeny, a taxonomy, or other information on history and relationships (such as archaeology or historical texts). Here are three alternative solutions to Galton's problem that use different forms of information on relationships between languages to deal with covariation due to descent. To avoid being out of date by the time this goes to print, I will describe the general approaches, rather than provide details on specific methods or programs, and readers are urged to find recent published studies in order to survey possible useful analysis methods and statistical packages (Harmon 2019; Höhna et al. 2016; Orme et al. 2013).

### *3.1 Solution 1: Use a phylogeny*

A phylogeny is a representation of the history of descent from a single ancestor to a number of descendants, usually represented as a series of nested bifurcating splits. While any means can be used to infer this series of splits, Bayesian phylogenetic methods are increasingly being applied to generate language phylogenies (e.g. Bouckaert et al. 2012; Bouckaert et al. 2018), as they allow the application of relatively rich models of change and provide a handy means of expressing uncertainty (Greenhill and Gray 2009). A language phylogeny represents the process of diversification of the languages themselves (e.g. Kitchen et al. 2009) but language phylogenies are also often interpreted as representing the history of human populations, tracking their fates and fortunes through time just as a molecular phylogeny based on human gene sequences might do (e.g. Gray et al. 2009; Grollemund et al. 2015; Lee and Hasegawa 2011).

There are two broad ways that phylogenies can be used to investigate correlation among features in language evolution. One is to infer a history of change, in order to examine the pattern or order of acquisition of language features (Levinson and Gray 2012). For example, Haynie and Bowern (2016) used a phylogeny of Australian languages to compare trajectories of colour term change. They tested the hypothesis that there is a predictable series of acquisition of colour terms as general terms are subdivided into more specific terms, starting with a black/white distinction, then adding red to the repertoire, then yellow and/or green, before adding blue and so on (Berlin and Kay 1969). Using a Bayesian analysis to model the gain and loss of colour terms along a phylogeny of 189 Pama-Nyungan languages, they supported an order-dependent model of gain of colour terms (certain colour terms are only gained once a particular colour term is already in the language), but also found support for other patterns such as loss of colour terms in some groups (Haynie and Bowern 2016).

The other way to use phylogenies to investigate correlation amongst language features is to ask whether two or more features co-occur together more often than expected by chance. This cannot be tested by simply tallying the number of times the features co-occur in the same languages, because even if two features have no functional connection they will tend to co-occur in related languages if they are inherited from a common ancestor that had those features (Figure 1). For example, it has been suggested that having a distinct future tense influences planning behaviour, such that people that speak languages that make a grammatical distinction between the present and the future are less inclined to save money for future use (Chen 2013). But people from related cultures might inherit both their language structure and their economic tendencies from their shared common ancestor: while both may change over time, it is possible that many populations inherit both future tense and low impetus for saving, so that these two traits co-occur more often than would be expected from a random distribution of those traits among cultural groups.

To investigate whether this shared inheritance could account for the association between future tense and saving, Roberts et al. (2015) used a Phylogenetic Generalised Least Squares (PGLS) regression. Ordinary least squares regression finds the straight line that minimizes the distance between each observed value and the line, in both the x and y dimension. But, as Galton pointed out, when our observed values are from relatives that have inherited their features from a common ancestor, then we ought to weight those observations accordingly (Galton 1889a). PGLS provides a principled way of weighting observations from different languages, by using a phylogeny to derive the expected correlation due to shared ancestry. PGLS has an advantage over other phylogenetic comparative methods that it provides a way of estimating the amount of covariance due to ancestry: if relatedness is not shaping variation in the data then the results of PGLS will be the same as an ordinary least squares regression (Symonds and Blomberg 2014). A PGLS analysis of future tense and savings supported the correlation overall, though it did not find the pattern within any of the language families that were analysed separately (Roberts et al 2015).

While phylogenetic analyses are commonly seen as the gold standard for dealing with Galton's problem, it is important to recognise that there are problems and limitations with this approach. A phylogeny is not a magical solution to the special problem of analysing data from evolved entities. Consideration must be given to the degree to which your data can discriminate alternative hypotheses. For example, if there are few instances of the origin of a particular language feature, then the data may have poor explanatory leverage for linking the origin of that feature to other language features, or for testing the effect of that feature on other aspects of language, due to lack of replication. It is not the number of languages you include in the analysis that count, but the number of independent occurrences of features. A p-value lower than 0.05 may guarantee publication, but it does not in and of itself guarantee that a meaningful relationship has been identified (Bromham et al. 2020). Phylogenetic tests are a useful tool for identifying and evaluating potential causal links, but they are not fool-proof evidence for or against hypotheses.

Of course, one of the obvious practical limitations is that phylogenies are not available for all groups (see section 3.2 below). But even when a phylogeny is available, careful consideration must be given to what the phylogeny represents. Few if any phylogenies are an error-free representation of the history of a group of languages. There are several sources of uncertainty. One arises from the phylogeny reconstruction process itself, which typically gives relative levels of support to alternative phylogenies, rather than supporting a single unambiguous representation of history. Increasingly, phylogenetic analyses are applied to a set of alternative reasonable phylogenies, such as a sample from the posterior distribution of a Bayesian analysis (e.g. Dunn et al. 2011). While this is a clear improvement on using a single tree to represent history, it should

not be considered to indemnify the analysis against phylogenetic uncertainty. The set of trees from a single phylogenetic analysis are all produced by the same data and assumptions. What we would really like to know is how sure we can be that the phylogeny represents the true history of descent of those languages. Would a reasonable adjustment to our assumptions, or slightly different data, result in a different tree? This level of error is much more difficult to characterise, though it can be explored by repeating the analysis on phylogenies obtained by different methods using different data and assumptions (Bromham et al. 2018a).

More deeply, we can question how well any tree can represent language history. In many cases, we do not expect the history of a group of languages to consist of a set of nested, two-way splits, with ancestral speaker populations dividing again and again to produce modern diversity. For many analyses, we do not need the phylogeny to be a literal representation of history, but to provide a description of the expected patterns of covariation between observations. Any phylogeny is better than none as long as it provides a better prediction of patterns of covariation that the default position of assuming all languages are equally similar to all others (Levinson and Gray 2012; Paradis 2014). However, for some phylogenetic analyses, the tree is treated as a map of history, for example when constructing the timing or order of acquisition of language features. In such cases, it would be prudent to explore whether alternative plausible histories would lead to the same conclusions.

Both kinds of phylogenetic analyses – order of acquisition along lineages and co-occurrence of features in descendants – implicitly rest on assumptions about the state at the internal nodes of the tree, representing the common ancestors from which current diversity is derived. The form of these assumptions about ancestral traits vary between methods. In some cases, there might be empirical evidence to inform past states, such as ancient texts or archaeological artefacts, though it will rarely be the case that they can be used to confidently assign past states to every node (branching point) in the whole tree. Instead, the state at the internal nodes must be inferred using a model of change (some Bayesian methods can integrate over possible states, but this still depends on the priors and assumptions that go into the transition model). While the outcome might resemble the reconstruction of protolanguage using the comparative method of historical linguistics, the process is somewhat different. Protolanguage reconstruction calls upon language-wide patterns of state changes (such as systematic sound correspondences) and requires the inferred states to form a coherent whole, whereas most phylogenetic methods consider features independently and using simple statistical models of change, such as Brownian motion (where trait values wander up and down stochastically over the phylogeny).

The need to impute past states is obvious for phylogenetic analyses of state changes which rely on historical reconstructions transitions from one state to another (e.g. Jordan 2011). But it is also inherent in many phylogenetic comparative methods that take a "whole tree" approach, even when ancestral states are not the focus of the study. It could be argued that inference at internal nodes is a statistical step in the analysis and should not be interpreted as an estimate of the actual trait values in any real ancestor (Harvey and Purvis 1991). Nonetheless, any analysis that makes inference about states that cannot be directly observed relies on assumptions about the process of change, and most such models are relatively simple stochastic models (Bromham 2019). This may lead to error when the underlying processes are not random but directional, for example systematic sound change, or where traits of interest have been borrowed between languages.

Inference of ancestral states also places special importance on branch lengths of the phylogeny. The models of evolution used to infer past states are usually conditioned on opportunity for change, typically expressed in terms of the chance of change per unit time. Therefore the expected amount of change from one node (ancestor) in the phylogeny to another (descendant) is usually a function of how long the branch connecting the two nodes is. Any aspects of the phylogenetic analysis that influence the inferred branch lengths (such as calibration dates or models of change) could potentially impact on the conclusions of a comparative analysis.

An alternative approach is to adapt phylogenetic methods to non-tree-like representations such as a network (Bastide et al. 2018). A language network is a typically a representation of the patterns of similarity and difference between languages (Figure 3). That is, it displays languages by using edges to connect languages that show similarities for the variables analysed (Bryant et al. 2005). While non-hierarchical patterns of similarity are often interpreted in light of processes of borrowing or contact (reticulate evolution) (Nelson-Sathi et al. 2011), these are not the only processes that generate connections on networks. For example, a grouping in a network analysis that includes Mandarin, Yoruba and Fijian is unlikely to represent either inheritance from a shared ancestor nor a history of contact and borrowing between these languages, but may instead reflect some shared language structures that occur in many different lineages (Greenhill et al. 2010).

Finally, there is cause for concern when the phylogeny or network used in a comparative analysis is derived from same data that is being analysed (Harvey and Purvis 1991). Using a tree or network based on word list to estimate rates of change of different lexical categories may be conflicted if the slow changing lexicon is an important source of grouping within the phylogeny. Such circularity should be avoided where possible, and interrogated for possible effect where it is unavoidable.

### *3.2 Solution 2: Use taxonomic information*

Lack of a suitable phylogeny does not make the problem of phylogenetic non-independence go away (Mace and Pagel 1994). Failing to correct for relatedness because information on relatedness is not readily available leaves an analysis vulnerable to misleading results, just as failure to include covarying factors in an analysis generates vulnerability to misinterpreting a correlation as a causal relationship (Bromham et al. 2020; Roberts and Winters 2013). In the words of one of the pioneers of phylogenetic analyses: "Comparative biologists may understandably feel frustrated upon being told that they need to know the phylogenies of their groups… Nevertheless, efforts to cope with the effects of the phylogeny will have to be made…there is no doing it without taking them into account" (Felsenstein 1985). Here, it is emphasized that the problem that must be addressed is the effects of phylogeny on our analysis (i.e. the patterns created by a process of descent), which is present in our data whether or not we know what the phylogeny is.

Yet there are clear limitations in the application of phylogenetic methods to understanding language evolution. Most language families do not have a published phylogeny, those that exist are often disputed and relationships between families are widely regarded as insoluble. More broadly, in many cases it may be reasonably questioned how well a simple bifurcating tree represents the tangled web of history and relationships between languages and cultures, and the patterns of language change over time. This situation may make the linguist feel they should not be expected to behave like biologists, for whom phylogenies are plentiful. And yet it is worth remembering that phylogenetic comparative methods were developed in biology long before phylogenies became widely available, and the demonstrated need for phylogenies to allow evolutionary analysis was one of the motivations that drove researchers to develop and apply phylogenetics in biology. In fact, not only were phylogenetic comparative methods developed long before phylogenies became readily available, but many researchers doubted there would be many available in future either (Huey et al. 2019). Instead, many early comparative analyses in biology used whatever information on relatedness was available to make an "assumed phylogeny" (e.g. Purvis and Bromham 1997), or included all possible phylogenies in the analysis (Martins 1996). Assumed phylogenies often rely on information from taxonomies, which continues to play an important role in comparative analyses in evolutionary biology. In fact, many of the large-scale evolutionary analyses use "megaphylogenies" that rely on taxonomic information to place species for which there are no DNA sequences available (e.g.Jetz et al. 2012; Webb and Donoghue 2005; Zanne et al. 2014).

Taxonomic databases of languages are a great boon to those studying language evolution (Eberhard et al. 2019; Hammarström et al. 2019). Of course, no taxonomy will be agreed upon by

all researchers, and yet there is clearly valuable information on relationships encoded in language taxonomies. Even if the exact relationships between languages in any given group are disputed, we may feel confident that all members of a particular group share a more recent common ancestor than any of them does with a far-flung language group from the other side of the world, and that we expect, on the whole, patterns of similarity and difference to be reflected those groupings. Taxonomic structure provides a platform for predicting patterns of covariation, which is precisely what is needed when comparing observations from different languages. Taxonomies may at least partially avoid the problem of circularity inherent in using language phylogenies (arising from cases where the same data are used to construct the phylogeny and to test hypotheses upon it), given that taxonomies often represent a qualitative statement of likely relationships based on an amalgam of many different observations rather than a quantitative analysis of comparative lexical data.

Taxonomies such as those in Ethnologue (Eberhard et al. 2019) or Glottolog (Hammarström et al. 2019) can be rendered into a tree either by hand or using freely available tools (e.g. Dediu 2018; Greenhill 2018). This tree does not have to consist entirely of bifurcating splits, each of which proposes an ancestral language split into two descendants. Instead, groups of related languages can be represented as "polytomies", comb-like structures that propose a common ancestor of a group of languages without indicating which languages in the group are more closely related to each other. This hierarchy should not be interpreted as a phylogeny as such: it represents patterns of relatedness, not the paths of divergence that ancestral languages took to produce contemporary diversity. For example, the branches of a tree made from a taxonomy should not be interpreted as evolutionary time. This is particularly the case where a tree is constructed for a sample of taxa and the nodes in the phylogeny are separated by default, unitary branch lengths (Bromham et al. 2018b). So how can taxonomies be used for methods that use branch length to infer expected patterns of change? One method is to distribute the nodes defining different taxonomic groups evenly, or scale them to the number of languages in the group (e.g. Roberts et al. 2015). However, this can have the effect of giving a greater time depth to groups containing more languages, which would not be reasonable if one group is diversifying more rapidly than another. An alternative approach is to fix the time depth to an accepted age for the group (or to an arbitrary age shared by all equivalent groups). Branch lengths can be scaled to give a relative reflection of the amount of expected similarity by assuming a time depth for family age and then using the number of levels of classification within the family to distribute nodes along the branches (see Bromham et al. 2018 for details). Use of a taxonomic tree with scaled branch lengths is not ideal, but it's a lot better than not correcting for relatedness at all.

### *3.3 Solution 3: Sister pairs*

Galton's problem of non-independence is a consequence of descent with modification: relatives start with the same ancestral features which may become modified over time, so the values of those traits in descendants show some level of dependence on the value in the ancestor, and therefore an association with among descendants. This is an ineluctable complication of comparing evolved entities. But the process of descent also provides a simple solution to the problem of non-independence, which is surprisingly robust to many of the problems of more sophisticated phylogenetic comparative methods discussed above.

One way to get around the problem of statistical non-independence due to descent would be to set up a replicated experiment. For example, if you were interested in the effect of population size on the patterns and rates of language evolution, you could take a bunch of humans with a shared language, divide them up into populations of different sizes, and see how their language changes over time in each replicate population (Raviv et al. 2019). You would need to leave the experiment to run for long enough for enough changes to accumulate to make meaningful comparisons of rate possible. For many targets of investigation, such as rate of lexical turnover or acquisition of grammatical features, the ideal experiment would take rather longer than the typical academic funding cycle and would be unlikely to make it past a sane ethics committee. But happily the experiment has already been run many times, without the intervention of university committees or funding agencies.

The peopling of the Pacific represents one of the greatest ocean-going migration events in the history of humanity. Expert navigators piloted double-hulled canoes over vast distances to establish populations on distant islands. The islands of Eastern Polynesia were the last great frontier of human migration, settled largely in the past millennium. Colonists arrived on the islands with everything they needed to establish settlement, and although they were not wholly isolated from other island societies, their languages inevitably changed over time and came to characterise each island culture. The carrying capacity for each population was apparently determined by available area, as the speaker populations of each island bears a predictable relationship to island size (Bromham et al. 2015). This resembles our ideal experiment – take a group of people with a shared language and distribute them into separate populations of different sizes, come back in a thousand years and see which ones have changed the fastest. But how do we compare rates of change?

One approach to comparing rates and population sizes would be to estimate rates of change along the branches of a phylogeny. But, even if a phylogeny is available, it is not straightforward to use it to compare rates of change in different languages. Language features can be recorded for any

sufficiently well-documented languages, but a rate requires a comparison between languages over a known timeframe. Rates depend on the accumulation of changes, so a rate estimate is expected to increase in precision the more changes have occurred. However, eventually so many changes have accumulated that past states may have been erased, making rates of change hard to accurately estimate due to saturation. So the statistical error in rates is related to both the speed of change and the time depth, and uncertainty in rates is likely to be high at the "shallow end" (recently diverged languages) and the "deep end" (ancient divergences) (Lanfear et al. 2010; Welch and Waxman 2008). To allow for the effect of time depth on comparisons, you need a way to fix the time interval across any comparisons, but you can't use the same data that you want to analyse rates for (for example, if you want to estimate rates of change in vocabulary it is problematic to use phylogenies with branch lengths estimated from lexical turnover).

There is a much simpler method that overcomes most of these problems, and does not require a fully-resolved, dated phylogeny. A sister pairs approach mimics an experimental design. The aim is to identify pairs of languages that are each other's closest relative. Any differences between the pair of languages must have evolved since they split from their shared common ancestor, whenever that was, so they have had the same amount of time to accumulate differences. This allows the calculation of relative rates of change: which sister has accumulated more differences since their shared ancestor? The acquisition or loss of language features in either sister lineage can be inferred by comparison to "outgroups" (languages that are more distantly related to either member of the pair). Selection of appropriate outgroups is a balancing act: too closely related and there is a danger that they may be more closely related to one of the sister languages than the other, too distantly related and the state in that language will be uninformative for determining the ancestral state of the sister pair. A good rule of thumb is to select the closest certain outgroup: the nearest relative that you can guarantee is more distantly related to either sister than either is to each other.

For example, we compared basic vocabulary between sister pairs of Polynesian languages to see if they shared cognate terms, or had lost or gained new words in those lexical categories (Bromham et al. 2015). If a particular cognate was found in only one language of the sister pair, but also in other languages in the family, this was taken as evidence that the cognate had been present in the common ancestor but then lost in one of the sister languages (Figure 4). If there was a novel word in one sister language that had no cognate in the other sister or any other members of the family, this was taken as evidence that the new word had been gained in that sister language alone. Of course, it is possible to think of ways that these assumptions could be violated for specific lexical items, but, given enough items of vocabulary (and as long as borrowings are excluded from the analysis), this comparison of cognates should indicate relative rates of gain and loss in a pair of

languages. As long as each pair consists of two languages whose most recent common ancestor is not shared with any other such pair, and each language acquires changes independently of any other such population, then the differences in rate are statistically independent observations that can be combined together in an analysis. When this approach was applied to pairs of closely related Polynesian language, the larger population had a higher rate of gain of new words more often than would be expected by chance, but the smaller population typically had a higher rate of word loss (Bromham et al. 2015). As for any such analysis, care must be taken in extrapolating beyond the observed cases to general principles. For example, a similar analysis found an association between small population size and word loss in Indo-European languages (Greenhill et al. 2018), but not in Austronesian or Bantu languages, suggesting that the effect may not be universal, or may be overridden by other factors.

A sister pairs analysis is a conservative approach, because you can select just those pairs that you are sure are each other's closest relatives, based on archaeological or linguistic evidence, and ignore any languages for which the relationships are complex or uncertain. But it is also flexible with regard to information on relatedness: a phylogeny or taxonomy could be used to choose sister pairs, but it might equally involve written or oral history, archaeological evidence, or any other information that attests the history and relationships between languages. In other words, you can focus on what you do know, rather than what you don't know. Another advantage of the sister pairs approach is that the underlying assumptions are much less onerous than for more formal phylogenetic methods. It easier to compare rates because both members of a sister pair have had the same amount of time to acquire differences. For example, comparison between sister pairs of birds show that the sister species closer to the equator tended to have a greater rate of change in syllable diversity and song length (Weir and Wheatcroft 2010). And it doesn't rely on a specific model of language change, because it does not require model-based inference of ancestral states. The major disadvantage of the sister pairs approach is that it typically involves including only a fraction of the languages for which you have data, because some languages will not be included in the analysis because they are not pair of a suitable pair, and no comparisons between ancestral nodes are made (as they would be in a classic independent contrasts approach: Figure 2). This can lead to a misplaced feeling of sorrow at "throwing data away", but it comes with the assurance that the datapoints you have are sound evidence. While a sister-pairs analysis may have fewer apparent datapoints, every datapoint is well-attested and based on actual observed languages, not on model-based inference of ancestral states.

## 4. Practical solutions to the problem of proximity

There are several ways that spatial proximity generates statistical non-independence in comparative language data. One is a direct effect of languages in contact influencing each other. For example, grammatical similarities between middle Andean languages from the Aymaran and Quechuan language families have been attributed to convergence rather than inheritance of a common grammatical system: continued contact between speaker populations might have favoured harmonisation of grammar (Adelaar 2012). But proximity will also create statistical non-independence even in the absence of significant contact or borrowing, through shared environment and history. For example, well-known spatial patterns in language diversity mean that two neighbouring areas are more likely to have similar values for language diversity than either has with a distant area (which is why language diversity correlates with both parasite load and chocolate consumption). Therefore unrelated variables that both show spatial patterning will tend to generate correlations between trait values (because the northern European countries with high chocolate consumption have relatively low language diversity, and tropical areas with high language diversity also have high parasite loads (Bromham et al. 2018b).

Shared spatial patterns may be caused by a shared covariant. For example, both language diversity and pathogen diversity have a latitudinal gradient (Collard and Foley 2002; Guernier et al. 2004; Mace and Pagel 1995). This may because aspects of climate, such as temperature and precipitation and growing season, influence both biodiversity and languages (Hua et al. 2019) – for example, it has been suggested that the length of the growing season per year influences language diversity by allowing more, smaller self-sufficient cultural groups to be supported within a given area (Nettle 1998). But it is also possible to get clustering of traits simply through the process of language evolution. For example, an association has been noted between the occurrence of tonal languages and humid climates (Everett et al. 2016). But if tonal languages tend to give rise to other tonal languages (or, to put it another way, if a language is more likely to be tonal if it is derived from an ancestral tonal language than if it had no tonal ancestors), then tonality will cluster on the phylogeny, and tonal languages will also cluster in space (Figure 5). Languages in the same area will experience similar climates, so there is likely to be an association between climate and tonality even in the absence of a causal link between the two. This effect can be magnified by the uneven distribution of languages across the globe: there are more languages in humid areas, so we should expect that there are also more tonal languages in humid areas (Collins 2016).

The clustering of tonal languages could also generate associations with anything else that shows geographic clustering, for example human genetic variants. Tonal languages have been shown to be associated with particular variants of genes (Dediu and Ladd 2007), and this association, along

with other evidence, has been interpreted as evidence for an interplay between human genetic adaptation and language evolution. But, due to histories of migration, settlement and intermixing, human populations tend to show spatial patterning of genetic variants (which is why DNA analysis is so useful for tracking the history of human populations). So we should not be surprised to find spatial patterns of language variation that correspond to spatial patterns in genetic variation, even in the absence of a causal link between genes and language. For example, tonal languages are also associated with genetic variants in human mitochondrial DNA (Collins 2017), even though it would be difficult to think of a direct association between mitochondrial variants and language tonality (given that mitochondrial genes are associated with basic processes of energy metabolism). Comparison to expected patterns of genetic difference will help to distinguish a pattern due to simple diffusion (Dediu and Ladd 2007), but any locally selected variant might correlate with clustered language features, even if not directly functionally related. These patterns do not invalidate adaptive hypotheses of language change, but they require us to rule out more mundane patterns of co-variation before calling upon adaptive explanations. Here I discuss two different kinds of data that can be used to account for spatial proximity – gridded variables from map-based data and pairwise distances between locations – but these share the same fundamental logic of using the distance between observations to generate an expectation of their likely covariation.

### *4.1 Solution 1: Grid-based spatial data*

One solution to the problem of proximity creating patterns of covariation is to do as Galton (1889a) suggested, and to put the data on a map and ask whether the association between the two variables of interest is greater than you would expect given their spatial distribution. In some cases, spatial data can be drawn from distributions available for local or global maps, which can then be used to derive other factors of interest such as climate, biodiversity or human population density. Information on spatial patterns of language diversity be generated from georeferenced polygons of geographical ranges of languages, such that the diversity of languages can be expressed for equal-area grids by counting how many languages overlap each grid cell (Amano et al. 2014; Hua et al. 2019).

For example, a study of language diversity and biodiversity in New Guinea used distribution maps to generate species and language counts for equal-area grids across the whole island. This led to the surprising conclusion that there is a significant negative correlation between endangered languages and endangered mammal species, because areas with high language endangerment tended to have low mammal endangerment and vice versa (Turvey and Pettorelli 2014). Treating each grid cell as an independent data point is equivalent to saying that at every location on New Guinea, the number of endangered languages is free to respond to variation in the number of

endangered species, or vice versa (or that both are free to respond to changes in some other unmeasured causal factor). In other words, each grid cell is treated as an independent test of the association between endangered languages and endangered species. But there might be broadscale geographic patterns that run over many different grid cells, such that a single event or process has affected the values in multiple datapoints.

When you plot the data on a map, it is clear that there are two distinct patterns; there are more endangered mammals in the highlands, and there are more threatened languages on the lowlands of the northern coast (Figure 6). Perhaps the distinct patterns in languages and species are due to different processes acting in different regions, such as different patterns of impact of human migration in highlands and lowlands due to differences in social or political history, or particular climatic events that affected the regions differently. Or perhaps there is some common factor, such as the difference in speaker population sizes between the areas. It might be that there are more endangered languages in the lowlands than the highlands due to historically smaller speaker population sizes, for example due to the effects of malaria (Foley 2000), while the much higher population density of the highlands put more pressure on mammals species through hunting and forest clearance (Flannery 1995). Whatever the explanation of this pattern, the practical upshot is that if you sample the number of endangered mammals and languages for grid squares covering the country, every time you sample a point in the northern lowlands you will find it has fewer threatened mammals and more threatened languages, and every time you sample a point in the highlands you will find it has high incidence of threatened mammals but low numbers of threatened languages (Bromham 2017). If you plot all of these points on a graph, you will see a pattern: many points with high language threat have low mammal threat, and vice versa, but this may be reflecting a single instantiation of an increase in language threat or species endangerment that affected multiple neighbouring grid cells, not a repeated test of what happens to one variable when the other changes.

You can deal with this pseudoreplication by taking the covariation between grid cells into account. The distance between grid cells can be used to generate a matrix of covariation that tells you how similar you should expect observations to be based only on their proximity. You can then look for any patterns above and beyond these expected levels of covariation due to distance. When you apply this analysis to the New Guinea data, there is no relationship between endangered languages and endangered mammals beyond what you would expect from their distribution on the map (Cardillo et al. 2015).

### *4.2 Solution 2: Distance between languages*

Cross-linguistic analyses often use variables collected at the level of languages or countries, with values for defined areas rather than continuously distributed. Relatedness and proximity are just as much of a problem for these analyses: neighbouring countries will tend to be more similar in many respects than they are to more distant countries, and this will generate associations between variables even if they have no causal connections. There are several ways to account for the proximity between languages, cultures or countries without requiring continuous spatial data. One approach is to use location as a factor in the analysis – that is, you add co-ordinates or other spatial information to your analysis along with all the variables you are interested in and see how much of the variation it explains (e.g. Dediu and Ladd 2007). This approach asks how much of the variation between observations can be explained by knowing their location. But a more principled approach is to use the location data as *a priori* information on expected patterns of covariation. This could be in the form of a matrix of pairwise distances, using, for example, the distance between the central point of countries, or between the borders of the language distributions, or whatever measure of distance is most salient to the question at hand. Using this approach, we start with the knowledge that some languages are closer to each other than others and that this is likely to shape the variation we see, then we ask if we can detect associations between variables above and beyond the association you get "for free" simply through the patterns of spatial proximity. This approach to accounting for covariation due to distance has the advantage that it can be dovetailed with the expected covariation due to relatedness, so that the expected variance is described using terms for both the phylogenetic distance and the physical distance between observations, plus a term that describes the relative contributions of relatedness and proximity (Freckleton and Jetz 2008). This approach allows comparison of the relative amount of influence of space and phylogeny on patterns of variation in the data (Hua et al. 2019).

**Conclusions**

> " There cannot be 'solutions' to the problems posed by comparative data, then, only approximation to solutions based upon our current understanding… the choice of a particular approach for analysing comparative data will often depend less upon knowledge that one technique is superior to another, than on a set of beliefs about the workings of evolution for a particular set of species and variables. "
> *Harvey PH, Pagel M (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford, page 120*

Comparative linguists have long been aware of the special challenges associated with analysing the products of descent with modification. The moniker 'Galton's problem' is usually taken to

refer to the problem of phylogenetic non-independence: close relatives are likely to be similar in many ways, and this can generate patterns of covariation between variables even in the absence of a direct causal connection. But Galton also highlight the problem of spatial non-independence: languages distributed in the same area share a history and environment that can generate apparent links between environmental features and aspects of language which may be due to shared distribution rather than a causal relationship.

Despite the long awareness of Galton's problem, there has not been a consistent approach to solving this problem that is universally applied or widely accepted. In part, this is due to a perception that it is not possible to correct for phylogenetic non-independence without a known phylogeny, which is a luxury most linguists don't have. But there are many ways of improving comparative analyses using whatever information on relatedness and proximity is available, for example from taxonomies, language compendia, archaeological evidence, ethnographic histories or sociopolitical information. The techniques described in this paper can be applied to any comparative study of languages, with the only limitation being whether the researcher can generate sufficient statistically independent observations from their data to make a valid test of alternative hypotheses. None of these approaches are perfect. But they are all a lot better than doing nothing to address Galton's problem.

**Acknowledgements:** I thank Marcel Cardillo for practical statistics help and valuable feedback, Andrew Ritchie for their helpful comments on the manuscript, and the MacroEvoEco group at ANU for enlightening discussions, including Xia Hua, Russell Dinnage, Hedvig Skirgård and Alex Skeels.

**References cited**

Adelaar WF (2012) Modeling convergence: Towards a reconstruction of the history of Quechuan–Aymaran interaction. Lingua 122(5):461-469

Amano T, Sandel B, Eager H, Bulteau E, Svenning J-C, Dalsgaard B, Rahbek C, Davies RG, Sutherland WJ (2014) Global distribution and drivers of language extinction risk. Proceedings of the Royal Society B: Biological Sciences 281(1793):20141574

Atkinson QD, Gray RD (2005) Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. Systematic Biology 54(4):513-526

Bastide P, Solís-Lemus C, Kriebel R, William Sparks K, Ané C (2018) Phylogenetic comparative methods on phylogenetic networks with reticulations. Systematic Biology 67(5):800-820

Berlin B, Kay P (1969) Basic Color Terms: Their Universality and Evolution. Univ of California Press, Berkeley, CA

Blomberg SP, Garland T, Jr., Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57(4):717-745

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD (2012) Mapping the origins and expansion of the Indo-European language family. Science 337(6097):957-960

Bouckaert RR, Bowern C, Atkinson QD (2018) The origin and expansion of Pama–Nyungan languages across Australia. Nature Ecology & Evolution 2(4):741-749

Bromham L (2016) Testing hypotheses in macroevolution. Studies in History and Philosophy of Science Part A 55:47-59

Bromham L (2017) Curiously the same: swapping tools between linguistics and evolutionary biology. Biology & Philosophy 32(6):855–886

Bromham L (2019) Six impossible things before breakfast: assumptions, models, and belief in molecular dating. Trends in Ecology & Evolution 34(5):474-486

Bromham L, Duchêne S, Hua X, Ritchie A, Duchêne D, Ho S (2018a) Bayesian molecular dating: Opening up the black box. Biological Reviews 93(2):1165-1191

Bromham L, Hua X, Cardillo M, Schneemann H, Greenhill SJ (2018b) Parasites and politics: why cross-cultural studies must control for relatedness, proximity and covariation. Royal Society Open Science 5(8):181100

Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ (2015) Rate of language evolution is affected by population size. Proceedings of the national Academy of Sciences USA 112: 2097–2102

Bromham L, Skeels A, Schneemann H, Dinnage R, Hua X (2020) Why do hot countries have spicy food? The challenges of testing hypotheses with cross-cultural data. Nature Human Behaviour Submitted

Bryant D, Filimon F, Gray R (2005) Untangling our past: Languages, trees, splits and networks. In: Mace R, Holden S, Shennan S (eds) The evolution of cultural diversity: a phylogenetic approach. Left Coast Press, Walnut Creek

Bulmer M (2003) Francis Galton: pioneer of heredity and biometry. The Johns Hopkins Press, Baltimore

Cardillo M, Bromham L, Greenhill SJ (2015) Links between language diversity and species richness can be confounded by spatial autocorrelation. Proceedings of the Royal Society B: Biological Sciences 282:20142986

Chen MK (2013) The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. American Economic Review 103(2):690-731

Collard IF, Foley RA (2002) Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? Evolutionary Ecology Research 4(3):371-383

Collins J (2016) Commentary: the role of language contact in creating correlations between humidity and tone. Journal of Language Evolution 1(1):46-52

Collins J (2017) Real and spurious correlations involving tonal languages. In: Enfield NJ (ed) Dependencies in language. Language Science Press., Berlin, pp 129–140

Daneyko T, Bentz C (2019) Click languages tend to have large consonant inventories: Implications for language evolution and change. In: Sahle Y, Reyes-Centeno H, Bentz C (eds) Modern Human Origins and Dispersal. Kerns Verlag,

Dediu D (2018) Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. Language Dynamics and Change 8(1):1-21

Dediu D, Ladd DR (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 104(26):10944-10949

Denis D (2001) The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists. History and Philosophy of Psychology Bulletin 13(2):36-44

Dow MM, Eff EA (2008) Global, regional, and local network autocorrelation in the standard cross-cultural sample. Cross-Cultural Research 42(2):148-171

Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473(7345):79-82

Dunn RR, Davies TJ, Harris NC, Gavin MC (2010) Global drivers of human pathogen richness and prevalence. Proceedings of the Royal Society B: Biological Sciences 277(1694):2587-2595

Eberhard DM, Simons GF, Fennig CD (2019) Ethnologue: Languages of the World. Twenty-second edition (www.ethnologue.com). In: SIL International, Dallas, Texas

Eff EA (2004) Does Mr. Galton still have a problem? Autocorrelation in the standard cross-cultural sample. World Cultures 15(2):153-170

Everett C, Blasí DE, Roberts SG (2016) Language evolution and climate: the case of desiccation and tone. Journal of Language Evolution 1(1):33-46

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1-15

Fincher CL, Thornhill R (2008) A parasite-driven wedge: infectious diseases may explain language and other biodiversity. Oikos 117(9):1289-1297

Flannery T (1995) Mammals of New Guinea. Reed Books, Sydney

Foley WA (2000) The Languages of New Guinea. Annual Review of Anthropology 29(1):357-404

Folyovich A, Jarecsny T, Janoska D, Dudas E, Beres-Molnar KA, Botos N, Biczo D, Toldi G (2019) Csokoládéfogyasztás és a magyar Nobel-díjasok [Chocolate consumption and Hungarian Nobel laureates]. Orv Hetil 160(1):26-29

Freckleton RP, Jetz W (2008) Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. Proceedings of the Royal Society B: Biological Sciences 276(1654):21-30

Galton F (1889a) Comment on 'On a Method of Investigating the Development of Institutions; Applied to Laws of Marriage and Descent' by E. B. Tylor. The Journal of the Anthropological Institute of Great Britain and Ireland 18:245-272

Galton F (1889b) I. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London 45(273-279):135-145

Galton F (1906) Cutting a round cake on scientific principles. Nature 75:173

Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323(5913):479-483

Greenhill SJ (2018) treemaker: A Python tool for constructing a Newick formatted tree from a set of classifications. The Journal of Open Source Software (JOSS) 3(31)

Greenhill SJ, Atkinson QD, Meade A, Gray RD (2010) The shape and tempo of language evolution. Proceedings of the Royal Society B: Biological Sciences 277(1693):2443-2450

Greenhill SJ, Blust R, Gray RD (2008) The Austronesian basic vocabulary database: from bioinformatics to lexomics. Evolutionary bioinformatics online 4:271

Greenhill SJ, Gray RD (2009) Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. Austronesian historical linguistics and culture history: a festschrift for Robert Blust. Canberra: Pacific Linguistics:375-397

Greenhill SJ, Hua X, Welsh CF, Schneemann H, Bromham L (2018) Population Size and the Rate of Language Evolution: A Test Across Indo-European, Austronesian, and Bantu Languages. Frontiers in Psychology 9:576

Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M (2015) Bantu expansion shows that habitat alters the route and pace of human dispersals. Proceedings of the National Academy of Sciences 112(43):13296-13301

Guernier V, Hochberg ME, Guégan J-F (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2(6):e141

Hammarström H, Forkel R, Haspelmath M (2019) Glottolog 4.0. (Available online at http://glottolog.org/) In: Max Planck Institute for the Science of Human History, Jena

Harmon LJ (2019) Phylogenetic Comparative Methods. Independent publication, Moscow, Idaho

Harvey PH, Pagel M (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Harvey PH, Purvis A (1991) Comparative methods for explaining adaptations. Nature 351(6328):619-624

Haynie HJ, Bowern C (2016) Phylogenetic approach to the evolution of color term systems. Proceedings of the National Academy of Sciences USA 113(48):13666-13671

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F (2016) RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65(4):726-736

Hua X, Greenhill SJ, Cardillo M, Schneemann H, Bromham L (2019) The ecological drivers of variation in global language diversity. Nature Communications 10(1):2047

Huey RB, Garland Jr T, Turelli M (2019) Revisiting a key innovation in evolutionary biology: Felsenstein's "Phylogenies and the comparative method". The American Naturalist 193(6):755-772

Jaeger TF, Graff P, Croft W, Pontillo D (2011) Mixed effect models for genetic and areal dependencies in linguistic typology. Linguistic Typology 15(2):281

Jetz W, Thomas G, Joy J, Hartmann K, Mooers A (2012) The global diversity of birds in space and time. Nature 491(7424):444-448

Jordan FM (2011) A phylogenetic analysis of the evolution of Austronesian sibling terminologies. Human biology 83(2):297-321

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. Proceedings of the Royal Society B: Biological Sciences 276(1668):2703-2710

Koplenig A (2019) Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. Royal Society open science 6(2):181274

Kummu M, Varis O (2011) The world by latitudes: A global analysis of human population, development level and environment across the north–south axis over the past half century. Applied geography 31(2):495-507

Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: Studying variation in rates of molecular evolution. Trends. Ecol. Evol. 25(9):495-503

Lee S, Hasegawa T (2011) Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Proceedings of the Royal Society B: Biological Sciences

Levinson SC, Gray RD (2012) Tools from evolutionary biology shed new light on the diversification of languages. Trends in cognitive sciences 16(3):167-173

Lupyan G, Dale R (2010) Language structure is partly determined by social structure. PLoS ONE 5(1):e8559

Mace R, Pagel M (1994) The comparative method in anthropology. Current Anthropology 35(5):549-564

Mace R, Pagel M (1995) A latitudinal gradient in the density of human languages in North America. Proceedings of the Royal Society of London B: Biological Sciences 261(1360):117-121

Maddieson I (2013a) Absence of Common Consonants. In: Dryer MS, Haspelmath M (eds) The world atlas of language structures online. The World Atlas of Language Structures Online. , Leipzig, p Available online at http://wals.info/

Maddieson I (2013b) Tone (https://wals.info/chapter/13). In: Dryer MS, Haspelmath M (eds) The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig

Martins EP (1996) Conducting Phylogenetic Comparative Studies When the Phylogeny is not Known. Evolution 50(1):12-22

Maurage P, Heeren A, Pesenti M (2013) Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies. The Journal of Nutrition 143(6):931-933

Messerli FH (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates. New England Journal of Medicine 367(16):1562-1564

Murdock GP, White DR (1969) A Standard Cross-Cultural Sample. Ethnology 9:329-369

Naroll R (1965) Galton's problem: The logic of cross-cultural analysis. Social Research:428-451

Nelson-Sathi S, List J-M, Geisler H, Fangerau H, Gray RD, Martin W (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. Proc Roy Soc B 278

Nettle D (1998) Explaining global patterns of language diversity. Journal of anthropological archaeology 17(4):354-374

Nunn CL, Altizer SM, Sechrest W, Cunningham AA (2005) Latitudinal gradients of parasite species richness in primates. Diversity and Distributions 11(3):249-256

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N (2013) The caper package: comparative analysis of phylogenetics and evolution in R. R package version 5(2):1-36

Paradis E (2014) An introduction to the phylogenetic comparative method. In: Modern phylogenetic comparative methods and their application in evolutionary biology. Springer, pp 3-18

Pearson K (1920) Notes on the History of Correlation. Biometrika 13(1):25-45

Poulin R (2014) Parasite biodiversity revisited: frontiers and constraints. . Int J Parasitology 44:581-589

Purvis A, Bromham L (1997) Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. Journal of Molecular Evolution 44(1):112-119

Rankin RL (2003) The comparative method. The handbook of historical linguistics:183-212

Raviv L, Meyer A, Lev-Ari S (2019) Larger communities create more systematic languages. Proceedings of the Royal Society B: Biological Sciences 286(1907):20191262

Roberts S, Winters J (2013) Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. PLoS ONE 8(8):e70902

Roberts SG, Winters J, Chen K (2015) Future Tense and Economic Decisions: Controlling for Cultural Evolution. PLOS ONE 10(7):e0132145

Shipley B (2002) Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press,

Sokolov AN, Pavlova MA, Klosterhalfen S, Enck P (2013) Chocolate and the brain: Neurobiological impact of cocoa flavanols on cognition and behavior. Neuroscience & Biobehavioral Reviews 37(10, Part 2):2445-2453

Stanton JM (2001) Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. Journal of Statistics Education 9:3

Symonds MRE, Blomberg SP (2014) A primer on phylogenetic least squares regression. In: Garamaszegi LZ (ed) Modern phylogenetic comparative methods and their application in evolutionary biology. Springer-Verlag, Berlin, pp 105-130

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL (2007) History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation. Molecular Biology and Evolution 24(10):2180-2195

Turvey ST, Pettorelli N (2014) Spatial congruence in language and species richness but not threat in the world's top linguistic hotspot. Proceedings of the Royal Society B: Biological Sciences 281(1796):20141644

Tylor EB (1889) On a Method of Investigating the Development of Institutions; Applied to Laws of Marriage and Descent. The Journal of the Anthropological Institute of Great Britain and Ireland 18:245-272

Webb CO, Donoghue MJ (2005) Phylomatic: tree assembly for applied phylogenetics. Molecular Ecology Notes 5(1):181-183

Weir JT, Wheatcroft D (2010) A latitudinal gradient in rates of evolution of avian syllable diversity and song length. Proceedings of the Royal Society B: Biological Sciences 278(1712):1713-1720

Welch JJ, Waxman D (2008) Calculating independent contrasts for the comparative study of substitution rates. J theor Biol 251:667-678

Willems M, Lord E, Laforest L, Labelle G, Lapointe F-J, Di Sciullo AM, Makarenkov V (2016) Using hybridization networks to retrace the evolution of Indo-European languages. BMC Evolutionary Biology 16(1):180

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB (2014) Three keys to the radiation of angiosperms into freezing environments. Nature 506(7486):89-92

**Figure 1:** Imagine a group of languages diversifying from a common ancestral language. Over time, one branch of the language group acquires two key changes (here represented by a thicker outline and a larger size) and the other branch acquires a different change (here represented by darker colour). If we treat contemporary languages as an association between size (x) and darker colour (y), then we would get a significant association even though the two features were acquired separately in different lineages. Similarly, if we looked for an association between outline and size, we would also find they tend to occur together, but this is based on one lineage that acquired both then produced many descendants. For example, we might see the same pattern if a group of languages with clicks (dark outline) evolved from an ancestor that had a large phoneme inventory (size), or if languages with larger average population sizes (size) happen to arise in a lineage that already has simpler grammar (darker outline).

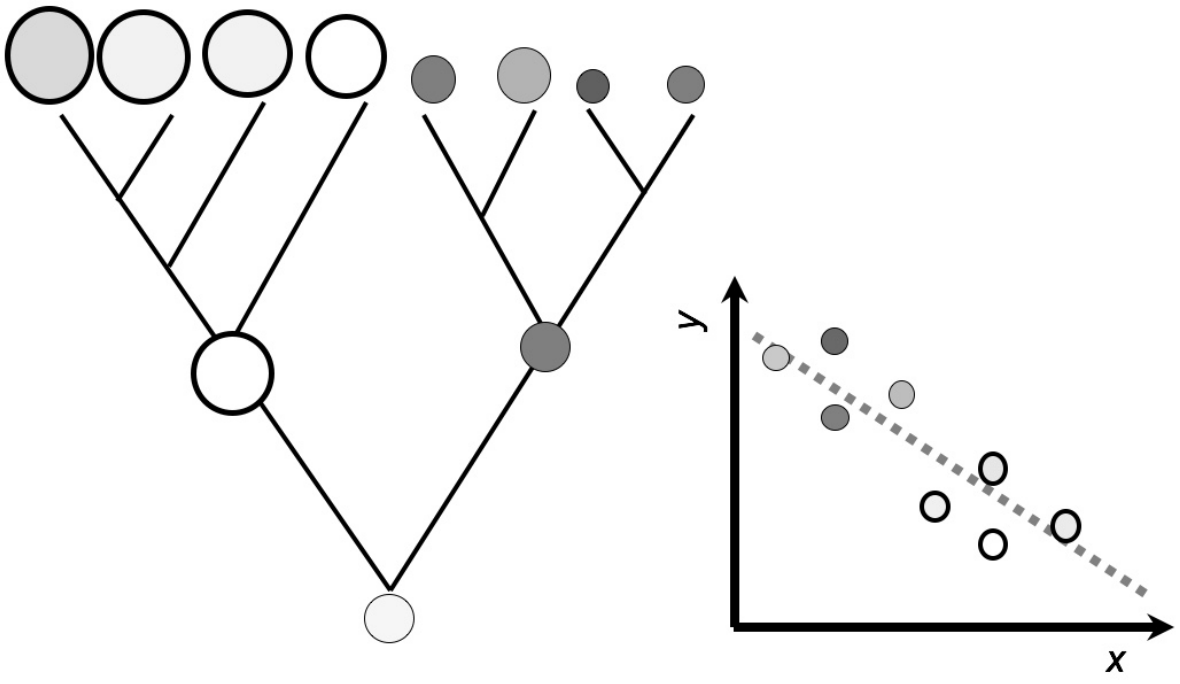
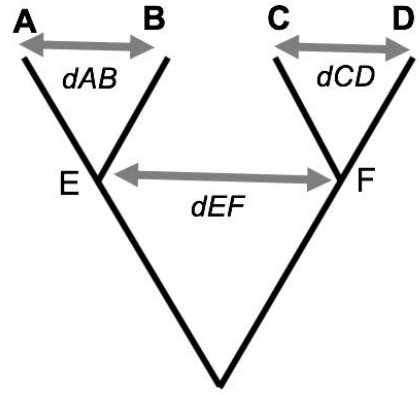


**Figure 2:** The method of phylogenetically independent contrasts (PIC) aims to select statistically independent contrasts, which represent the differences in variables ($x$, y) that pairs of languages have acquired since they both started from the same value in their common ancestor. The contrasts between the values for the members of each pair are the datapoints that are used in an analysis, such as a correlation. Observations on contemporary languages can be compared by identifying pairs of "tips" on a phylogeny (here, the comparisons between A and B, and between C and D), an approach known as sister pairs analysis. But if the ancestral states at the branching points (nodes) of the phylogeny are known, then independent contrasts can also be identified between past languages (here, E and F). It might be possible to infer those states using documentary information such as ancient texts or historical records. But more usually the state at the internal nodes (E and F) is inferred using a model of evolution to identify the most likely state, given the state in the descendant nodes (A, B, C, D) combined with a model of change. As with all such methods, the reliability of the results depends on the veracity of the observed tip states (languages A to D) and on the evolutionary model being an adequate representation of the processes of change. This figure is based on an illustration from Harvey and Pagel (1991).

| *language* | A | B | C | D |
|---|---|---|---|---|
| *y* | 7 | 9 | 14 | 20 |
| *x* | 20 | 24 | 30 | 40 |



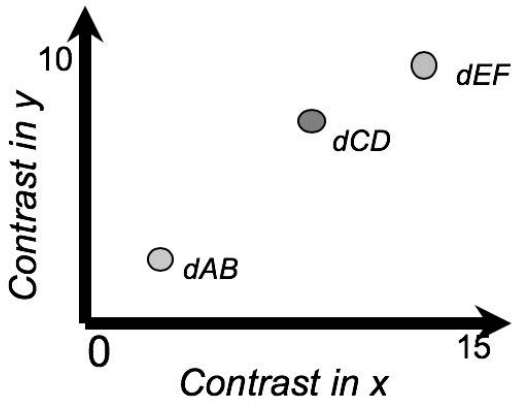



**Figure 3: Example of phylogenetic network and phylogeny** for 8 languages of the West-Germanic group from (Willems et al. 2016). The lines connecting lineages in the network (a) represent the relative proportions of shared features of languages. Shared features might arise from inheritance, borrowing, or independent acquisition. The same pattern of shared features can be used to construct a phylogeny (b) which show the most strongly supported groups as a series of nested bifurcating splits. For languages that have not evolved strictly through a series of two-way splits, there may be "non-tree-like" patterns in the data, here represented as the "boxes" in the network (a) or dotted lines across the phylogeny (b). For example, Sranan is a creole incorporating lexicon from both English and Dutch. Figure reproduced with modification under Creative Commons Attribution 4.0 International licence from Willems M, Lord E, Laforest L, Labelle G, Lapointe F-J, Di Sciullo AM, Makarenkov V (2016) Using hybridization networks to retrace the evolution of Indo-European languages. BMC Evolutionary Biology 16(1):180

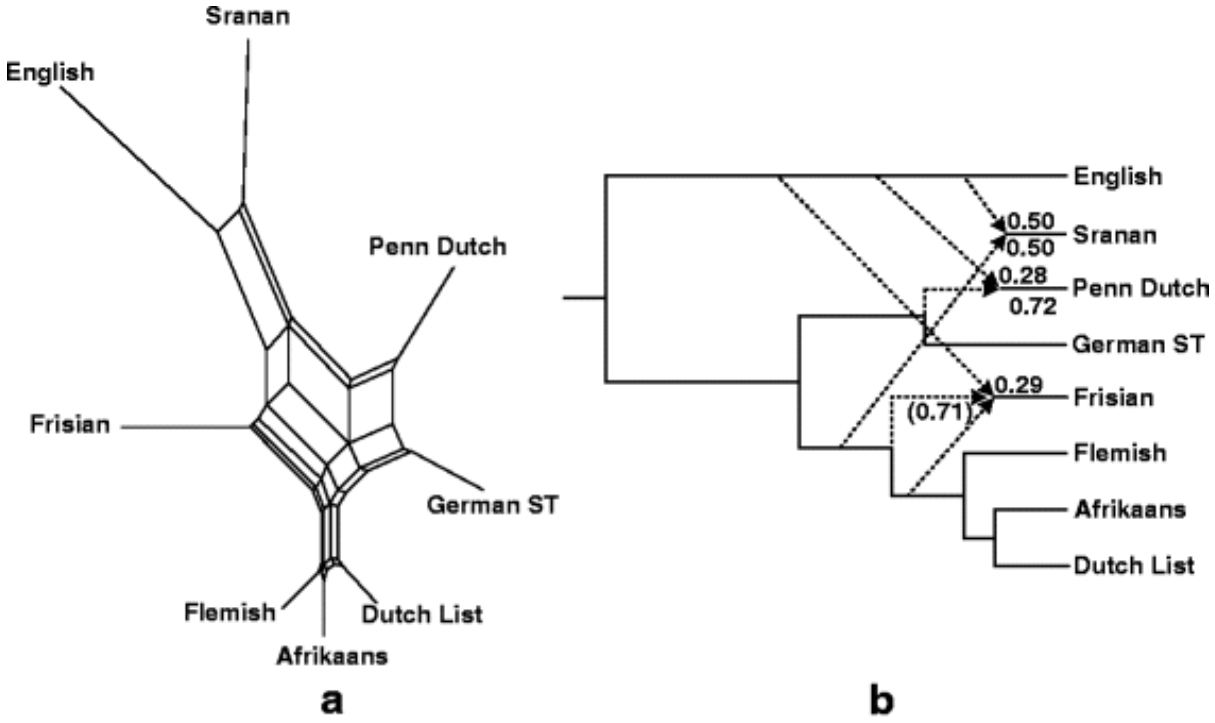


**Figure 4:** A comparative test for different rates of gain and loss of language features in a sister pair of languages. This test was adapted from a test for difference in rates of DNA sequence evolution in different lineages (Tajima 1993). Note that no phylogenetic information is needed apart from identifying a pair of languages that are each other's closest relatives, and information on the state of a language variable in both members of the pair and some other members of the language group. The group can be any set of related families that are all considered to share a common ancestor. This test was first used to compare rates of gain and loss of words from basic vocabulary in Polynesian languages (Bromham et al. 2015): the shared forms were cognates identified from the Austronesian Basic Vocabulary Database (Greenhill et al. 2008), the language group was considered to be the Austronesian family of languages, and pairs were identified from a range of sources including linguistics and archaeology. However the basic principle would apply to any language features where it is possible to identify sisters pairs of languages which have shared forms due to common inheritance and unique forms that have arisen independently. The language group could be any grouping that identifies a set of languages with a common ancestor, such that the sister pairs are all nested within that group, and no pair overlaps with any other pair (that is, the most recent common ancestor of each pair is not an ancestor of any other pairs in the analysis). This test relies on the assumption that borrowings are not included and cognates share a single common origin, so that the presence of a cognate in more than one language is evidence of shared inheritance. Based on a figure first published in Bromham et al. (2015).

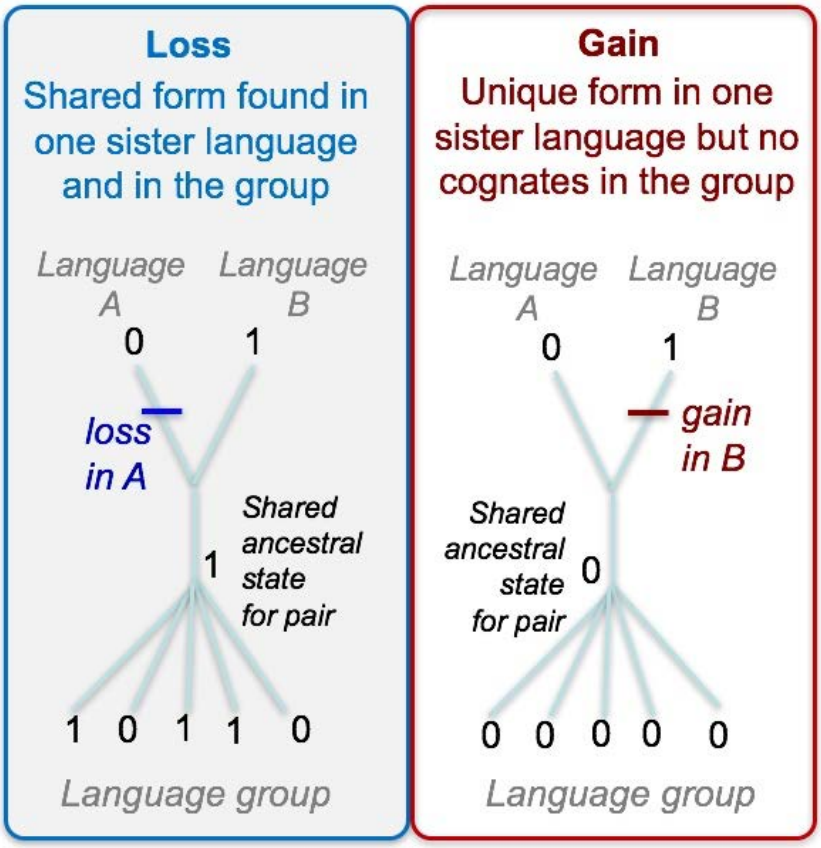


**Figure 5.** Languages with simple tones (pink) or complex tones (red) cluster in space. Image from the WALS database (Maddieson 2013b). Map is produced by OpenStreetMap (© OpenStreetMap contributors) and is reproduced under Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA 2.0).

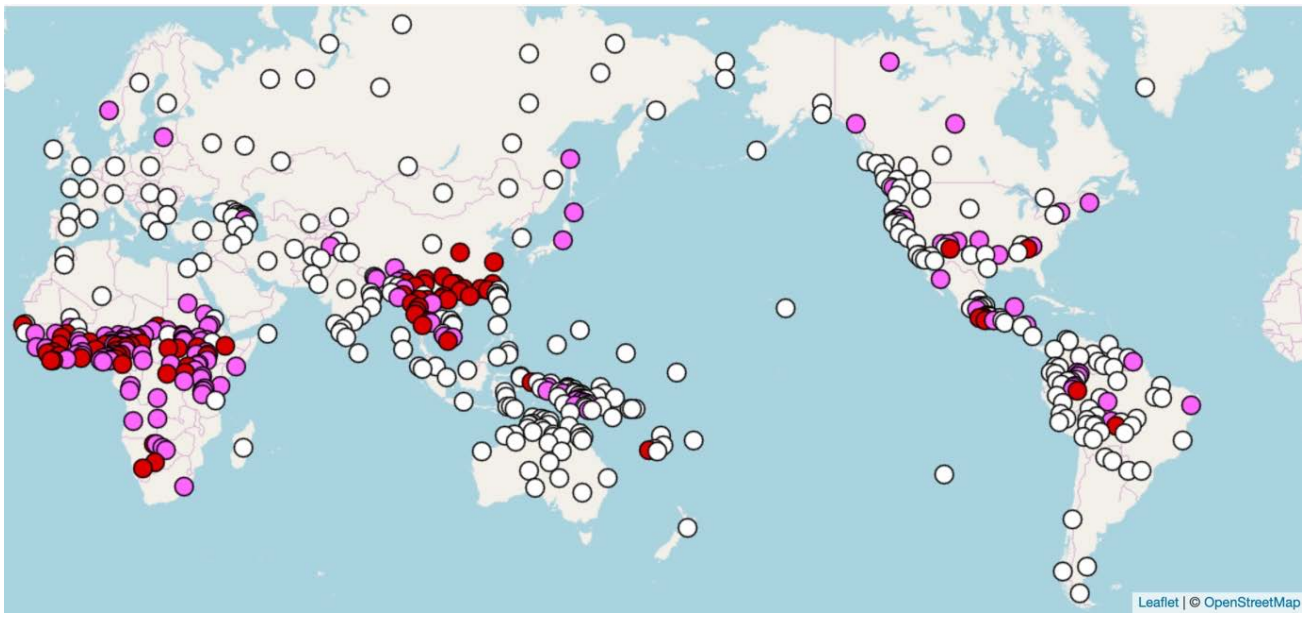


**Figure 6:** Distribution of languages and mammal species in New Guinea. The high elevation areas (a) have high mammal diversity (b), whereas language diversity tends to be greater on the northern lowlands (c). The number of threatened mammal species matches the areas of higher diversity: where there are more mammal species, at areas of high elevation, there tend to be more endangered mammal species (d). Endangered languages show the opposite pattern, being generally lower in the areas of high elevation (e). When you plot both threatened mammal species (f: hatched areas) and threatened languages (f: shaded areas) on the same map, they are largely non-overlapping patterns, meaning that most areas with high threatened mammals tend to have low threatened languages. Correcting for spatial autocorrelation suggests there is no significant relationship between threatened mammals and threatened languages beyond that caused by these spatial patterns. Modified from Turvey & Pettorelli (2014) and reproduced under Creative Commons license CC-BY version 4.0