# Language phylogenies

Michael Dunn
Max Planck Institute for Psycholingusitics

12 August 2013

## 1 Introduction

In principle, all historical linguistics is phylogenetic, since phylogenetics encompasses the scientific investigation of the descent of organisms in general. While prototypical phylogenetic analysis involves investigating the evolutionary descent of a class of biological species, such phylogenetic analyses have also been applied in other domains (e.g. social organization, musical instruments, decorative motifs on textiles); and indeed, can be applied to any domain which varies according to general evolutionary processes. The reason that in linguistics this term is often used in contrast to other forms of historical linguistic investigation is that phylogenetic approaches maintain their methodological link to the investigation of evolutionary processes in other, mostly biological, domains. The appeal of incorporating the analysis of language into a general theory of evolution is that current evolutionary theory offers a rigorous, quantifiable approach to phylogenetic inference. As Felsenstein states in the preface to his monumental *Inferring Phylogenies*, "phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyse those differences statistically" (Felsenstein 2004: xix). Linguistic phylogenetics incorporates the whole approach of the phylogenetic comparative method — using language phylogenies as the historical backbone to quantitative models of language change in order to test hypotheses about human dispersals, processes of cultural change, and the evolution of other linguistic subsystems (see Section 5). In this sense phylogenetic linguistics is broader in its ambitions than historical linguistics: historical linguistics seeks to illuminate the history of languages, and only secondarily seeks to say something about the speakers of those languages in approaches like *linguistic palaeontology* and *socio-cultural reconstrction* (see Chapters **XREF:Heggarty** and **XREF:Epps**), as well as *gene-language correlation* (Chapter **XREF:Packendorf**).

The quantified, algorithmic approach to phylogenetics started in the early 1960s (Felsenstein 2004). Linguistics has been part of this movement twice: firstly with the development of *lexicostatistics* and *glottochronology* in the late 1960s, and again with the development of model-based, hypothesis-testing (and usually *Bayesian*) approaches starting around 2000 (see Section 5.3). The aim of this chapter is to present a overview of current quantitative approaches to language change. These approaches are increasingly well received within linguistics, but remain controversial in some circles, in part because of the conflation of lexicostatistical approaches

with better theoretically grounded approaches used today, and partially because of a mismatch between the aims and scope of the linguistic Comparative Method and the statistical, hypothesis testing approach of quantitative methods.

## 2 Inferring linguistic phylogenies

Most phylogenetic analyses seeking to infer language history are based on modelling the historical behaviour of lexical cognate sets, typically as represented in a Swadesh list or other standardised list of meanings. In its raw form, a Swadesh list contains a set of lexemes corresponding to meanings. In a well-made Swadesh list the criteria for assigning particular lexemes to these meanings are properly defined. According to Swadesh himself, the lexeme should be the stylistically unmarked, everyday word corresponding to the meaning. According to Swadesh, "[t]he rules for filling in the list for each language may be stated as follows: a) Try to find one simple equivalent for each item by disregarding specialized and bound forms and the less common of two equivalents. b) Use a single word or element rather than a phrase, even though the meaning may be broader than that of the test item. c) Where it is impossible to find a single equivalent, omit the form." (Swadesh 1952). Obviously, care should be taken that words correspond to the intended meaning (e.g. Swadesh' "bark" is the "skin of a tree", not the "noise of a dog"). Kassian et al. (2010) have produced a useful semantic specification for a widely used version of the Swadesh lists. In a phylogenetic analysis the lexemes are replaced by their cognate class, rather than dealing with the lexical forms themselves, so that all lexemes which descend from a common ancestor are indicated by the same code.

| | Meaning 1 | | Meaning 2 | |
|---|---|---|---|---|
| | lexeme | class | lexeme | class |
| Language A | mhɨm | a | cɨŋ | x |
| Language B | mhɨm | a | kɨt | y |
| Language C | lɔːt | b | kət, lpəc | y, z |
| Language D | ? | ? | lpət | z |

Table 1: Multistate coding of two meanings; a, b, x, y, z are cognate classes; "?" means unknown

Table 1 shows an example fragment of a cognate coding matrix from a Swadesh list. Technically, this is a *multistate* matrix, since each meaning (a *character* in phylogenetic terms) may have multiple values (e.g. *x*, *y* or *z*). This matrix can be transformed into a binary presence-absence matrix by treating each cognate class as a character; see Table 2. The presence-absence matrix has the useful property that it can handle more than one cognate set per meaning (as in Language C, Meaning 2).

Correct cognate classification is no trivial matter. Dunn, Greenhill, et al. (2011) and Bouckaert et al. (2012) used a database of cognate classifications checked against published data from etymological dictionaries. Cognate judgements may also be supplied by historical linguists, as in Ringe, Warnow, and Taylor (2002), and Greenhill, Blust, and Gray (2008). Phylogenetic analyses can also be carried out on *cognate candidate* data, where cognate classification in the strict sense (proved by historical linguistic methods) has not been (or possibly cannot be) carried out. The

| | Cognate set | | | | |
|---|---|---|---|---|---|
| | 1a | 1b | 2x | 2y | 2z |
| Language A | 1 | 0 | 1 | 0 | 0 |
| Language B | 1 | 0 | 0 | 1 | 0 |
| Language C | 0 | 1 | 0 | 1 | 1 |
| Language D | ? | ? | 0 | 0 | 1 |

Table 2: Binary (presence-absence) coding of cognate data from Table 1

data in these two cases looks identical, but using cognate candidates rather than proven cognates necessarily adds a further (unfortunately unquantified) level of uncertainty to the analysis. Coding for lexical candidates rather than true lexical cognates is crucial for other kinds of analysis, especially those where the research question is specifically about borrowing or admixture. For instance, Shijulal et al. (2011) model horizontal transfer of lexical items inside families, and Bowern (2012) uses cognate candidates to classify 19th century wordlists of Tasmanian languages into probable languages and language families.

In principle, anything which carries a phylogenetic signal can be used as the basis for some kind of phylogenetic inference, although of course the kind of phylogenetic signal constrains the inferences which can be made. As in conventional historical linguistics, phylogenetic analysis can model notionally unique events in the history of a language family such as sound changes and morphological innovations. The *perfect phylogenies* approach (Ringe, Warnow, and Taylor 2002; Nakhleh, Ringe, and Warnow 2005) most closely approximates a fusion of traditional and phylogenetic comparative methods. Their database (Nakhleh et al. 2005) includes cognate sets, as well as features representing morphological and phonological innovations; their statistical analyses are designed to incorporate the same assumptions that a traditional historical linguistic analysis would make, seeking a tree which is maximally compatible with all the evidence for subgrouping (potentially including evidence which is conflicting in the perfect phylogenetic networks approach; Nakhleh, Ringe, and Warnow 2005).

Other kinds of analyses are not so tightly coupled to traditional historical linguistics as those based on cognates and sound-changes. If traditional historical linguistic input is unavailable or for some other reason unfeasible, then approaches to phylogenetic analysis based on phonological similarity are appealing. These methods use some kind of formally defined distance measure to compare lexemes between languages. This measure can be used to identify cognate candidates (Dunn and Terrill 2012), but more often it is used to give an estimate of the amount of evolutionary change between lexemes which are a priori presumed to be cognate. Lexical similarity methods are particularly suited to investigation of dialect data, since the presumption of cognacy is well justified (Prokić 2010). In general, lexical similarity measures work best at relatively shallow time-depths (Greenhill 2011).

Phylogenetic methods are not limited to working with lexical, morphological or phonological features. In work on the historical connections between the Papuan languages of Island Melanesia, for example, Dunn et al. (2005) investigated the hypothesis that the typological similarities between these languages retains a historical signal of earlier contact or shared ancestry,

despite the absence (or just unavailability) of lexical evidence. Because of the limited "design space" of language (Reesink and Dunn 2012), the probability of chance similarity in the typological domain is high, and particular care must be taken to use appropriate statistical tests if such data is to be used to support detailed claims about history, or to make inferences about deep time depths. Models of structural typological data presume that unrelated languages may have identical values for some typological parameters, which make them particularly useful for the analysis of contact and admixture (Reesink, Singer, and Dunn 2009), and gene-language correlation (Hunley et al. 2008).

## 3 Distance-based models of change

A distance-based model of change estimates the amount of change between two languages from the aggregate amount of difference between them. Distance-based methods of phylogenetic inference all use some kind of *distance metric* to measure how much each taxon differs from every other one. There are two different kinds of distance metric which are commonly used in linguistics: shared cognate proportion and so-called Levenshtein distance, a word-by-word measure of phonetic similarity.

### 3.1 Lexicostatistics

Language clustering by cognate distance is the earliest form of statistical phylogenetics done in linguistics. A standardized list of meanings (a "Swadesh list" or the like; Section 2) is used to compile wordlists representing each language of the sample. For each pair of languages, the distance between them is the proportion of corresponding terms which are not cognate in the two lists (The inverse of distance, 1 - d, is the proportion of terms which *are* cognate between the two lists, and can be thought of as the similarity or proximity of the lists). The pairwise distances between the languages are tabulated, as in Table 3.

| | Language A | Language B | Language C | Language D |
|---|---|---|---|---|
| Language A | - | | | |
| Language B | 2/5 = 0.4 | - | | |
| Language C | 5/5 = 1.0 | 3/5 = 0.6 | - | |
| Language D | 3/3 = 1.0 | 2/3 = 0.66 | 1/3 = 0.33 | - |

Table 3: Pairwise distances from Table 2. Language A and Language B differ in 2 our of 5 comparisons, so have a distance of 0.4; because of missing data ('?') distances from Language D are calculated on the basis of only 3 comparisons.

It is important that the criteria for inclusion in these lists are clear, otherwise biases to the distance calculation may be introduced simply due to availability of materials for a particular language. According to the Swadesh 1952 criteria cited above, there is minimal allowance made for semantic change (known cognate terms which don't fit the semantic criteria must be ignored). Swadesh also does not allow cells to contain multiple lexemes (as in Language C, Meaning 2 in Table 1). But in each of these cases this is an analytic choice: other criteria are possible, but the distance measurements are only coherent if these criteria are applied consistently.

The term glottochronology is sometimes used as a synonym of lexicostatistics, but some

scholars are punctilious about distinguishing the terms. If the terms are to be distinguished, lexicostatistics refers to the process of clustering languages based on distances calculated from meaning lists, whereas glottochronology refers to a method for using these distances to infer chronological dates (McMahon and McMahon 2005: 33). The glottochronological method for determining dates is based on the premise that over a sufficiently large tree, is is acceptable to assume a constant rate of change. In the late 1940 and 1950s the promise of glottochronology seemed vast: Swadesh (1952) talks breathlessly about the "discovery" of a mean rate constant of lexical cognate turnover, believed to be 81% ± 2 per 1000 years for culturally neutral vocabulary. According to this constant, a pair of languages which split 1000 years ago would share 81% of the terms in a meaning list; a 2000 year old split would show 81% × 81% = 66% retention, and so on. If you accept (both the validity and the value of) this rate constant it is fairly simple to determine the age of any language relationship (Swadesh 1952: 461–462). The problem with this whole line of research is that the constant rate of change turned out to be illusory. The rate of language change varies due to a number of different factors. Amongst those to attempt to quantify this is Nettle (1999), who shows a general pattern of faster language change in smaller communities (the same phenomenon as faster genetic drift in smaller biological populations), and Atkinson et al. (2008), who show that higher rates of cognate turnover are associated with language splitting events. Computer simulations by computational biologists have shown that distance-based clustering is extremely sensitive to differences in rate of change in different branches of the tree (Peer 2009: 147 for references), and other, more robust methods now exist (see Section 4). A role remains for distance methods: the Automated Similarity Judgement Program (ASJP; Brown et al. 2008; Holman et al. 2011) makes extensive use of distances (but see Greenhill 2011), as do the dialectometric techniques discussed in Section 3.3.

### 3.2 Levenshtein distance

Instead of working from a distance measure based on cognate classification, it is also possible to work directly from the similarity between the phonological forms of words. This is usually done using some variant of the Levenshtein distance, a measure of how many operations are required to turn one string of phonemes into another. To change a "hawk" into a "handsaw" requires two substitutions and three insertions, for a distance of five (it can also be done with one deletion and four insertions; see Figure 1 for another illustration). The distances between one short word and one long word will always be great (because of the number of deletions required), and so it is common to normalize the measure along a 0-to-1 scale by dividing by the length of the longer word. The Levenshtein distance measure can be customized in many other ways too. Different weights can be given to insertions and deletions, and the distance cost of a substitution can be calculated from phonological features (so to change a /t/ to a /d/ counts as less distance than a change from a /t/ to a /ŋ/). The substitution measure can also be made more coarse, so that classes of similar phonemes are treated as identical, as is done in the ASJP project (Holman et al. 2011). In the ASJP all phonemes are collapsed into 41 classes (so, for example, L represents all laterals except for /l/; Brown et al. 2008) to allow maximal cross-linguistic comparability. The loss of phonological distinctiveness is compensated for by the extremely large scope of the

comparison — 4817 languages and dialects at the time of the Holman et al. 2011 paper.

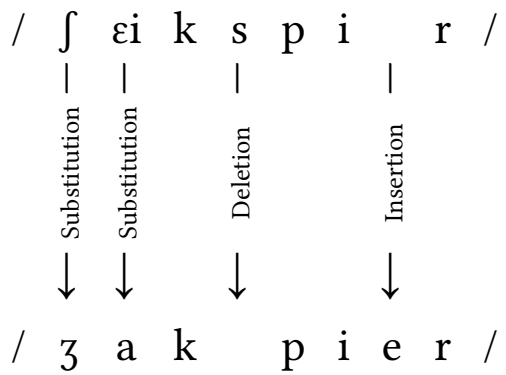


Figure 1: Levenshtein distance: How to turn a "Shakespeare" into "Jacques Pierre" (in broad phonological transcription) with two substitutions, a deletion, and an insertion, for a Levenshtein distance of four.

### 3.3 When to use distance, and when not to use it

One serious criticism of the Levenshtein distance measure is that it is only a coherent measure of language change where the forms being compared descend from a common ancestor. The Levenshtein distance between non-cognate items is not predicted by the degree of relatedness of the two languages that the words come from. In the *computational dialectometry* tradition (Goebl 1993; Heeringa and Nerbonne 2001) this objection is largely irrelevant, since most words in a comparative dialectology wordlist *are* cognate. In addition, the time depths of separation between dialectal varieties is generally much smaller than between varieties recognised as distinct languages, meaning that variation in rates of change has less influence on the amount of linguistic difference. Taking these two factors into account, the amount of phonological difference between two dialects can be considered a good proxy for their amount of historical separation. Dialect diversification tends to be wavelike, rather than treelike, and qualitative differences between dialects are usually represented by isoglosses. These quantitative approaches to dialectology differ from the isogloss approach in that isoglosses show discontinuities, while distance measures give a continuous clustering. Computational dialectometry measures are usually analysed and visualized using network methods. One vivid technique for visualizing dialect clusters is to transform the pairwise Levenshtein distance matrix (i.e. the distance calculated from each variety to every other variety) using multidimensional scaling (MDS) with three dimensions. These dimensions can then be converted to intensities of the three colour axes in the Red-Green-Blue colour space, and the resulting colours can then be used to colour the dialect polygons on a geographic map (Heeringa 2004). In general, varieties with similar relationships of similarity and difference to the other varieties considered will have similar MDS values, which in turn will produce visually similar colours. Note that the Red-Green-Blue colour space is not perceptually even: humans discriminate many more hues in some areas of the colour space than others, meaning that visualization by RGB transformed MDS values is intrinsically distorted. Real, scientific analysis of distance relationships must be numeric. For example, Manni,

Guérard, and Heyer (2004) use Monmonier's algorithm to detect *barriers*, abrupt changes in linguistic distance uncorrelated with geographic space.

Levenshtein distance is a good proxy, however, for historical relatedness when the languages being compared share a recent common ancestor. Because Levenshtein distance between non-cognate lexemes is not a meaningful measure of historical relatedness, this measure is a particularly risky technique to use for long distance comparison. It becomes increasingly unreliable when trying to quantify older linguistic relationships. If a pair of languages are not related at all, the Levenshtein distance between them is solely a function of their similarities in phonology and phonotactics, as well as chance. Dunn and Terrill (2012) present a method for investigating the base rate of chance similarity between words from two different languages, based on the "Oswalt Shift Test" (Oswalt 1970, 1998). The Oswalt Shift Test was conceived of as a validity test for lexicostatistics: the rate of shared cognate candidates between two lists was compared with the rates of shared cognates between lists with their rows offset by one, two, three, etc. Thus, for a list of 100 meanings you would be able to compare the percentage of apparent shared cognates in the true, semantically aligned list to the percentages of apparent cognates in 99 semantically unaligned lists. This test would obviously be quite difficult to do by hand, since the linguist would have to make 1000 cognate judgements knowing that all but one percent of them were bogus comparisons. It would be desirable to do the cognate coding blind, with no knowledge about either of the languages involved which, in practice, would mean that this painfully tedious academic task could only be done by somebody with no prior exposure to the area. Unsurprisingly, the Oswalt Test was rarely carried out in practice. Dunn and Terrill (2012) propose a modernization of the Oswalt Shift Test, called the Oswalt Monte Carlo Test. This is a version of the Shift Test using a Levenshtein measure to propose cognate candidates. The distance threshold and the parameters of the precise variant of the Levenshtein measure are tuned against a set of training data. This makes it practicable to carry out large numbers of comparisons, so the ad hoc "shift" method for selecting 99 semantically unaligned lists to compare can be replaced by the statistically more standard Monte Carlo randomization technique, where the wordlists are randomly shuffled thousands of times in order to get a reliable sample of the distribution of distance measures under the semantically scrambled condition. ASJP comparisons have begun to use an Oswalt Monte Carlo-type procedure to correct their mean Levenshtein distance by dividing it by the distances between scrambled wordlists. In principle this should indicate that the mean Levenshtein distance measured between one pair of languages should be comparable to the mean Levenshtein distance measured between any other pair of languages. One outcome of the case study of presented in Dunn and Terrill (2012) was to demonstrate how vulnerable distance measures are to undetected loanwords. An apparent high degree of lexical commonality between the non-Austronesian languages of the Solomon Islands vanishes when probable loanwords are removed. Even one or two loanwords in a wordlist give a strong false-positive signal.

## 4 Character-based models of change

An alternative family of approaches to phylogenetic inference, known as *character-based models of change*, estimate the relationship between two languages by inferring the pathways by which each evolved from their common ancestor. The difference between two languages according to a distance model is always equal to or less than the difference according to a character-based model. Another way of looking at this is that the distance between two languages is the shortest path from one to another. A character-based model is constrained to provide an evolutionarily plausible pathway to each language from their inferred common ancestor---and this is almost always a longer pathway than the shortest distance for turning one language into another. The shortcuts that distance measures take are more serious the further back in time the common ancestor is located. Character-based methods are thus more realistic models of evolutionary processes.

The earliest, and until recently the most widely used character-based method for inferring phylogeny was the parsimony method (Swofford and Sullivan 2009:268). The parsimony method seeks a tree that explains a data set (e.g. a set of cognate judgements) by minimizing the number of evolutionary changes required to produce the observed states. This means that tree structures are preferred that place innovations where they account for the greatest amount of observed diversity as possible. This is a similar logic to that used used in the linguistic Comparative Method: trees are constructed on the basis of shared innovations, and where possible, a tree topology is found in which each innovation has occurred only once. Studies using simulated data show that parsimony methods are weak at recovering the true evolutionary tree in particular kinds of conditions. The most serious of these is probably *long branch attraction*: long branches (branches with a lot of change) in a tree will tend to be clustered together even if they are only distantly related in the true evolutionary history. This occurs because where two branches have both undergone a lot of change, the most parsimonious account is always to bundle the two branches together as a single set of innovations with a relatively small amount of independent diversification at the end. Parsimony methods have recently been overtaken by statistically more robust, but computationally harder, *likelihood* methods, which are not subject to this problem.

### 4.1 Likelihood methods

Likelihood methods seek to explain a set of observed data by quantifying how likely it was to have been produced by a particular process. The likelihood (L) is the probability of seeing the observed data (D) under a particular hypothetical mechanism (H), formalized as L = P(D|H). Within phylogenetics, the hypothesised mechanism is an evolutionary process, referred to as the "model", which consists of a mathematical description of evolutionary change. A model might typically include tree topology and branch lengths, as well as, e.g., the probability that a new cognate set appears in the tree, and the probability that a reflex of a cognate set is lost. The model itself is not in question within the likelihood calculation: the model is the researcher's hypothesis about the mechanisms of evolutionary history (but different models can be compared and evaluated, see Section 4.3.1).

Maximizing the likelihood of the parameters of a tree (the most likely branch lengths, the

most likely transition parameter values, etc.) is generally tractable to exact mathematical methods. However finding the best tree topology out of the vast space of possible trees is extremely challenging, and no algorithm is known that guarantees that this best tree will be found in any reasonable amount of computational time (Schmidt and Haeseler 2009). It is not possible to solve this using random sampling of tree likelihoods: since they are very skewed towards low likelihood values, and only a tiny proportion of the trees in the space represent good solutions to the evolutionary hypothesis. Most reasonably sized random samples of trees from this space will contain no high likelihood trees at all, and there would be no way to know that the highest likelihood trees in the sample would count as high within the distributions of likelihoods in the entire space. The practicable solution to this is to use Bayesian Monte Carlo Markov chain (MCMC) sampling. This algorithm searches the tree space for the region of highest likelihood. It starts at a random point in the space, and by randomly perturbing the parameter values at that space, compares the likelihood score of the current position in the tree space to the new set of values. If the new values have higher likelihood, then the same search is repeated from the new position. This functions like a simple hill-climbing algorithm: if likelihood is elevation, then the search reaches out and measures the height difference of a nearby point, and if it is higher, takes a step in that direction. In this overly simple version of the algorithm the search would risk getting stuck at a local maximum (at the top of a foothill, rather than at the top of the mountain); the real algorithm has a number of ways for dealing with this. Rather than completely ignoring lower likelihood positions in the tree space, the search will randomly accept proposals to move to lower likelihood positions in proportion to differences in likelihood (so it would be much more likely to accept a move to a position which is only slightly lower, but would most likely reject a move to a very much lower likelihood position). Every thousand steps or so (to avoid *autocorrelation*, sampling of statistically non-independent trees from too close in the parameter space), the tree parameters and likelihood value are saved to the *posterior sample*. Figure 2 shows a typical record of likelihood (or elevation, for hill climbing) over the search. If all goes well, after an initial period of wild oscillation, the search reaches a reasonably level, maximum state. The initial period — the "burn-in" — is discarded, and the remainder of the sample represents more-or-less equally highly likely sets of parameters (strictly speaking, they are sampled in proportion to their likelihood). In phylogenetic inference these parameters typically include tree topology, branch lengths, and transition probabilities.

### 4.2 Evolutionary models

In Section 3.1 it was mentioned that one of the great theoretical weaknesses of glottochronology was that it assumed a constant rate of change. Rate of change is included in likelihood analyses as the *clock model*. A *strict clock* corresponds to the constant rate of change assumption, and in biology as in linguistics, it has been shown empirically to be inappropriate in many cases (Felsenstein 2004: 322–329). Several other more realistic clock models have been devised. So called *relaxed clock* methods relax the strict clock assumption by allowing rates to vary across the tree, chosen from a probability distribution whose mean is determined by the rate of the parent branch (Drummond et al. 2006). This relaxed clock parameter is incorporated as part

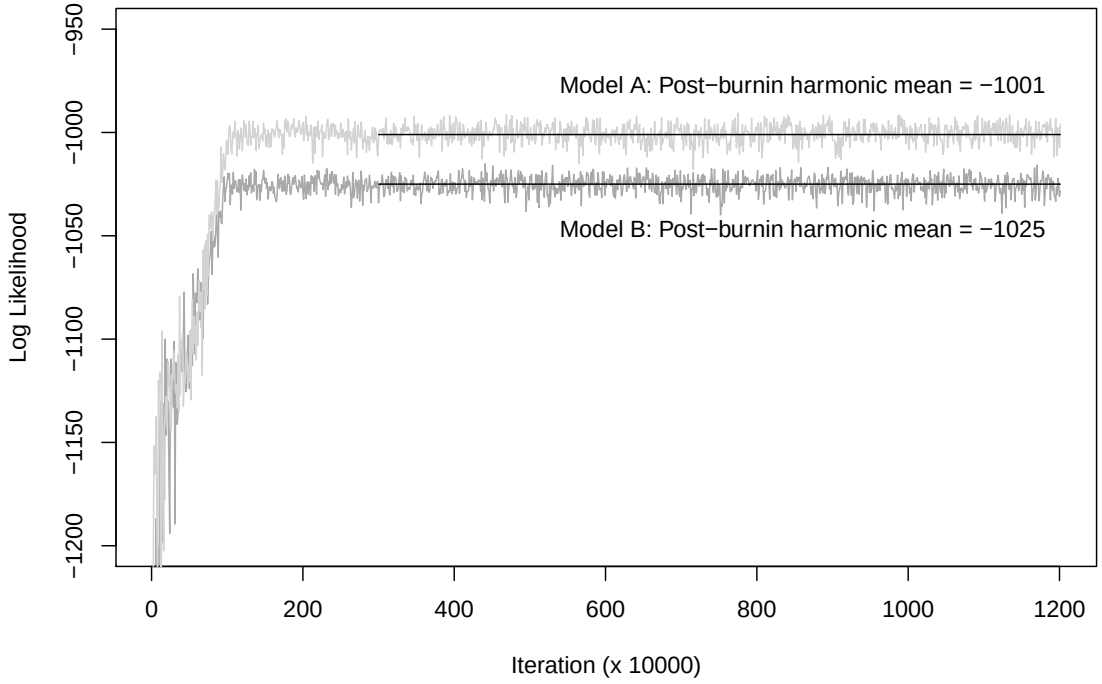


Figure 2: Likelihood trace MCMC sampling under two models, A and B. The Bayes Factor = 2 × (likelihood of Model A - likelihood of Model B) = 48, which indicates "very strong" support for Model A (see Section 4.3.1).

of the a priori model of evolutionary change. Different kinds of probability distribution can model processes where rate change occurs continuously along a branch, or where rates change at nodes independently of branch length. Another clock model, the *random local clocks* model (Drummond and Suchard 2010), treats rate variation as a series of independent local clocks, each extending over a subregion of the complete tree. The number of distinct local clocks required to account for the observed data is a parameter estimated by this model, which makes this technique applicable to questions of clade-specific rate variation. A note should also be made here of *no clock* models. In cases where it is unclear which clock model is appropriate for a data set, it might be tempting to use a model which enforces no clock at all. But this would be a mistake: a no clocks model explicitly states that there is no limit on the evolutionary variation between branches — an assumption which is even less likely to be a good fit to the real evolution processes than any of the other clock models available (Pybus 2006).

Along with the clock model, which specifies how rates change globally across branches, a substitution model must also be specified, indicating how rates differ from character to character (e.g. from cognate set to cognate set). Model complexity is defined by the number of parameters the model has, and in general a good model should have as few parameters as possible (Section 4.3.1). The basic substitution models for binary data are the simple one-rate or two-rate models (Figure 4a). In a one-rate model, a single parameter $q$ expresses the probability that a character state 1 will turn into 0 or vice versa. In a two-rate model the probability that 1 will turn into 0 ($q10$) is estimated separately from the probability that 0 will turn into 1 ($q01$). This is

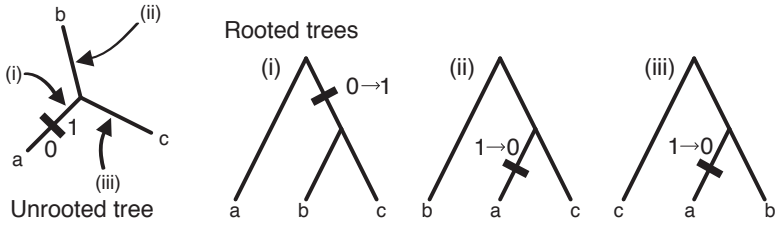


Figure 3: Unrooted and rooted trees. An unrooted tree makes no assumptions about chronology. Three possible roots are marked, (i–iii); the bar on the branch leading to *a* indicates a change between *state 0* and *state 1* of a feature known to occur on this branch of the tree. Each root hypothesis allows you to make different inferences about subgrouping and directions of change: If the tree is rooted on (i), the split between *a* and *(b,c)* is the earliest in the tree, and thus *b* and *c* are sisters and the evolutionary change of state is from *0* to *1*; if the tree is rooted on (ii) then *a* and *c* are sisters, and rooting on (iii) puts the earliest division between *c* and *(a,b)*; in both these latter cases the evolutionary change of the feature is *1* to *0*.

a more realistic model, in that the rate of innovation is treated separately from the rate of loss. A tree produced under a two rate model must be rooted, since you need to know the root position to know whether a change is $0 \longrightarrow 1$ or $1 \longrightarrow 0$ (see Figure 3). Some phylogenetics packages require that the root be specified by the researcher, but since the root selection determines the overall tree likelihood under any given rates hypothesis, it is possible to estimate the tree root as part of the topology (e.g. as in BEAST; Section 7). The *gamma model* assumes that each character belongs to one of a specified number of rate classes . Figure 4b schematically illustrates a three class, two rate version of this model. The rates for each class are sampled from a gamma distribution, a probability distribution with the useful property that its shape is controlled by a single parameter. This could be expected to be a better fit to linguistic data, as different terms on the Swadesh list have different stability. The *covarion model* (Figure 4c) gives another way to account for rate variation in the data. Under the covarion model the rate of each character is allowed to vary along the branches of the tree. The Stochastic Dollo model (Figure 4d) captures a key feature of real cognate histories, that a cognate set may only be innovated once in the dataset. One parameter governs the distribution of rates of cognate sets arising in the data, and another parameter the distribution of rates of loss of their reflexes (Nicholls and Gray 2008). The Stochastic Dollo model necessarily produces a rooted tree.

Note that the Stochastic Dollo model would *not* be suitable for modelling the history of typological features. Typological features evolve in a more limited design space than lexical cognates, with a correspondingly higher probability of chance homology, and so a reversible model would be more appropriate (for this reason it is in any case most unlikely in the model selection procedure, outlined in Section 4.3.1, that the Stochastic Dollo model would out-perform Gamma or Covarion models on a structural database).

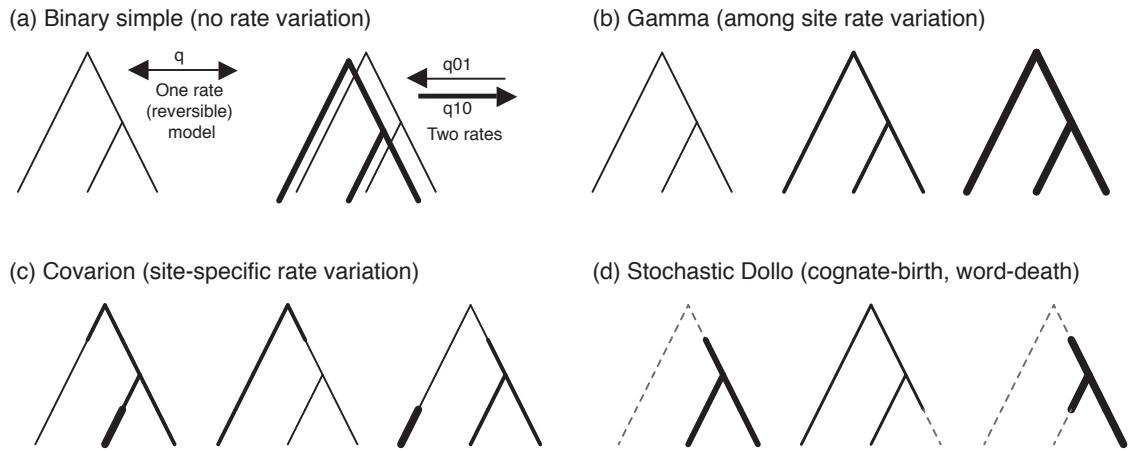


Figure 4: Substitution models. Gains and losses of reflexes of cognate sets are modelled as (a) occuring at a single rate, or at a distinct rate for gains and a for losses, (b) different rates for different cognate sets, (c) different rates for different branches, and (d) a rate of innovation and a rate of loss.

### 4.3 Interpreting the results

#### 4.3.1 Model choice

The likelihood score of an analysis is the probability for the observed data to evolve given a particular model. Even assuming that for each model the optimal parameter values have been inferred, some models still fit better than others. The difference in acceptability of two models can be expressed by the Bayes Factor, which is calculated as the ratio of $L(H_1)$, the likelihood of hypothesis one (also expressed as "$Pr(D|H_1)$", the probability of the data D under $H_1$) to the likelihood of $L(H_2)$ (or "$Pr(D|H_2)$"), as follows:

$$BF_{12} = \frac{L(H_1)}{L(H_2)}$$

The Bayes Factor statistic is often expressed as twice its natural logarithm, $2\log BF_{12}$, rather than using the raw ratio $BF_{12}$. Likelihood ratios are the same as the difference in log-likelihoods; thus $2\log BF_{12} = 2(\log L(H_1) - \log L(H_2))$. The interpretation of the Bayes Factor statistics are given in Table 4. The Bayes Factor is not like a p-value: it is only meaningful in testing one hypothesis against another. A negative Bayes Factor simply supports $H_2$ over $H_1$ (which incidentally means that the Bayes Factor test can give evidence in favour of the null hypothesis; the strongest evidence frequentist statistics can give for the null hypothesis is not to rule it out). Note also that the probability of getting the observed data with a model has nothing to do with the probability that the model is correct. The best model is as simple as possible, but not simpler (a maxim attributed to Einstein). Where likelihood ratios do not provide a clear front-runner, the analysis should be based upon the front-running model with the fewest parameters.

#### 4.3.2 Tree sample

The results of a Bayesian phylogenetic inference analysis are a sample of trees (as well as logs of other model parameters of interest associated with each of these trees), sampled in propor-

| $BF_{12}$ | $2\log BF_{12}$ | Evidence for $H_1$ over $H_2$ |
|---|---|---|
| 0 to 2 | 1 to 2 | Negligible |
| 3 to 20 | 2 to 6 | Positive |
| 20 to 50 | 6 to 10 | Strong |
| >150 | >10 | Very strong |

Table 4: Guidelines for the interpretation of Bayes Factors and Log Bayes Factors (after Kass and Raftery 1995: 777)

tion to their posterior probability. This sample typically runs to thousands of trees, and it is of course not feasible simply to inspect it by eye. Summarizing all this output in a single, coherent narrative is difficult. This is usually best done using some form of visual summary. There are several possibilities. For a tree sample, most practitioners currently prefer a maximum clade credibility tree (Figure 5a). This is a tree selected from the tree sample which maximizes the product of likelihoods of each of its branches. This tree can be treated as the best representative of the tree sample. Branch lengths are usually taken from the median or mean of corresponding branches in the sample. A maximum clade credibility tree can be constructed using the TreeAnnotator tool of the BEAST package (Drummond et al. 2012). The consensus tree method has also been popular as a way of summarizing sets of trees. Consensus trees are built by ranking all the binary splits in a tree sample and building a tree which includes all the branches that don't contradict any more highly ranked branch. One of the disadvantages of a consensus tree is that there is no guarantee that a tree with the same topology as the consensus tree will actually be present in the tree sample. There is no generally accepted way to specify the branch lengths of a consensus tree. Consensus trees can be made using many different software packages; the "consensus" function of the APE package in R (Paradis, Claude, and Strimmer 2004; see also Section 7) is a convenient and flexible option. Whichever kind of tree representation is used, it is normally desirable to annotate branches with confidence estimates (their posterior probabilities for a Bayesian tree; other kinds of confidence estimates such as bootstrap values exist for non-Bayesian trees). A DensiTree visualization of the tree sample gives a vivid representation of the entire tree sample by overplotting all the trees of the sample in partial transparency (Figure 5b). Finally, if conflicting groupings in the tree sample are of interest, the tree sample can be summarised with a *consensus network* (Holland et al. 2004). Each branch of a tree in the tree sample represents a possible binary split of the data. These splits are weighted according to their frequency, and used to compute a splits graph. Where the splits are compatible, the splits graph is identical to a tree; where splits are incompatible, each split is represented by a collection of parallel edges, where incompatible splits are orthogonal (see Figure 5c). The length of the edges can be used to represent the weight (in this case, frequency) of the split.

#### 4.3.3 Priors

One of the great advantages of the Bayesian approach to inference is that it allows you to integrate many different forms of prior knowledge. Apart from the distributional priors on model parameters (Section 4.2), it is frequently desirable to integrate a priori known elements of tree structure. This is done by constraining the MCMC search to those parts of the parameter space

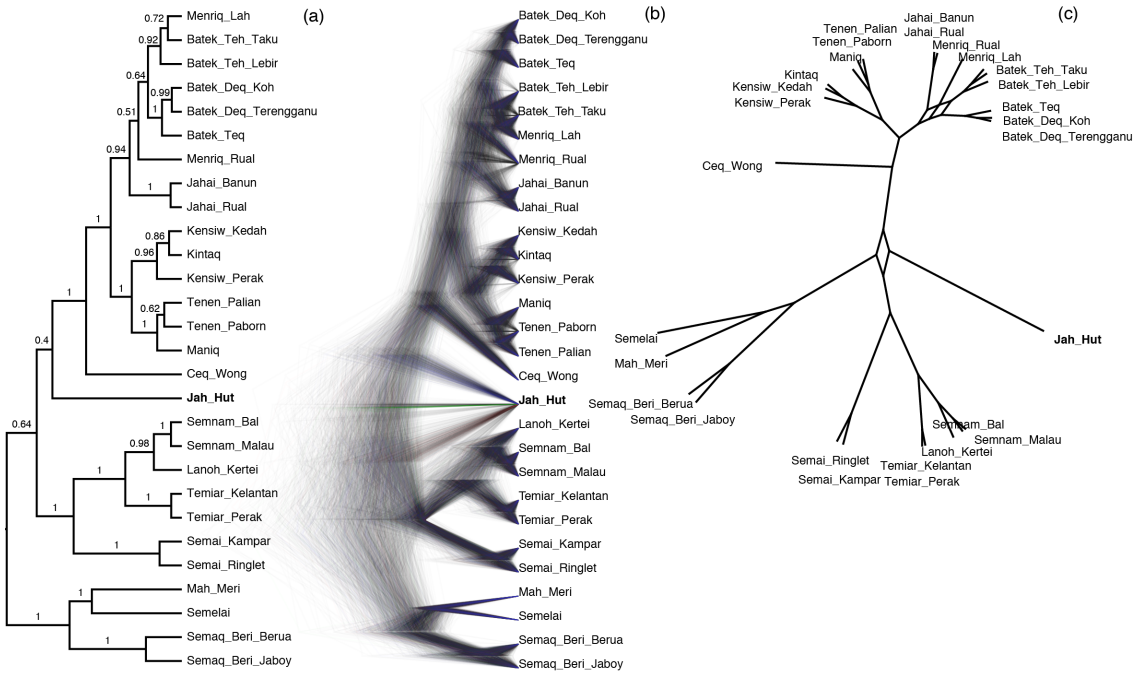


Figure 5: Summarizing the posterior tree sample; Aslian phylogenies (Dunn, Burenhult, et al. 2011) visualized with (a) Maximum Clade Credibility (MCC) Tree, (b) DensiTree, and (c) Consensus Network. Note how the uncertainty about the classification of the language "Jah Hut" is reflected by (a) low posterior probability values, (b) multiple points of origin, and (c) a box showing conflicting splits.

where the tree topologies are consistent with the prior knowledge. This can have the simple practical advantage of speeding up the analysis — if the data contain a strong signal for particular elements of structure, then searching other parts of the tree space may simply be unnecessary. But tree priors are also the most direct way to integrate subgrouping knowledge based on linguistic Comparative Method criteria, such as phonological and morphological innovations. Constraining the tree search to areas of the parameter space which are consistent with comparative method innovation-based subgrouping is a perfectly valid way of integrating traditional comparative method and computational phylogenetic methods of tree inference, and can give results which are greater than the sum of their parts: combining comparative method subgrouping information into a Bayesian MCMC analysis as priors allows inference of (i) trees which are further resolved than the subgroupings provided by sound changes; (ii) comparative method trees with meaningful branch lengths or chronological calibrations; and (iii) comparative method trees with quantified estimates of uncertainty and rate change.

In addition, Bayesian Phylogenetic inference methods can be used to test between competing sets of priors (using the likelihood ratios test; Section 4.3.1). For instance, where there are conflicting tree constraints based on apparent shared innovations a likelihood ratio test would allow a test of whether the cognate histories support one subgrouping hypothesis over the other.

### 4.4 More complex cases

It is also possible to use phonological and morphological innovations as characters in the analysis, rather than constraints. This is the approach taken by Ringe, Warnow, and Taylor (2002) and Nakhleh (2011), in an analysis which seeks to reconcile both cognate histories and shared phonological and morphological innovations. Within the Bayesian Phylogenetic Inference approach it would probably be most appropriate to do this using *partitions* — carrying out the analysis with a different model used for cognate evolution than for morphological/phonological innovations. A similar partitioned analysis can be used to carry out phylogenetic inference on other kinds of disparate data where different evolutionary processes apply to different parts of the data.

Dunn et al. (2008) and Dunn (2009) present Bayesian Phylogenetic inferences based on typological features of language rather than lexicon. The rationale in this case was to investigate the history of a set of languages showing typological similarities which were suggestive of a historical relationship, but which did not show lexical similarities sufficient to demonstrate their relatedness (See also Dunn and Terrill 2012). This approach has been adopted for languages where sufficiently complete word lists are unavailable (e.g. Daniëlsen, Dunn, and Muysken 2011). The reliability of historical inference based on typological features is the subject of some controversy (Donohue and Musgrave 2007; Dunn et al. 2007); a non-controversial use of typological features is for comparison of the linguistic similarity between languages which are not (or not known to be) related for the purposes of hypothesis generation or testing patterns of genetic versus linguistic diversity (Dunn 2009; Reesink, Singer, and Dunn 2009; Hunley et al. 2007, 2008).

A number of Bayesian phylogenetic methods deal with reticulation. The reticulated tree model of Shijulal et al. (2011) adapts some of the insights of evolutionary microbiology to linguistic questions of *family-internal borrowing*. Bacterial evolution is highly complex, as bacteria have a number of mechanisms which allow them to combine and exchange DNA, including swapping of the DNA coding for complete functional subsystems (e.g. antibiotic resistance). A naïve reconstruction of evolutionary history of bacteria would infer an ancestral mega-bacterium, already embodying the precursors to all the DNA diversity of its descendants. The essential insight of this method is to apply the uniformitarian hypothesis, that bacteria in the past were not substantially different to bacteria in the present; more precisely, the size of the genome of ancestral bacteria should fit within the distribution of observed genome sizes in present-day bacteria. A Bayesian phylogenetic tree search is carried out with an additional "horizontal gene transfer" parameter. This parameter is optimised to maintain the inferred genome size at ancestral nodes to within the statistical distribution of known bacteria. The linguistic adaptation of this uses the same logic: rather than inferring an ancestral language containing all the lexical diversity of the contemporary languages, the algorithm infers lexical transfer where it can efficiently account for shared cognates in distant branches. By keeping the inferred size of the ancestral lexicon within the bounds of the observed distribution, this avoids pushing the date of proto-forms further back than they need to be.

Ringe, Warnow, and Taylor (2002) take a different kind of perspective to the problem of

phylogenetic inference from the Bayesian inference approach. They develop an algorithm to test the perfect phylogeny problem, that is, whether a tree can be found which perfectly reconciles a set of phonological, morphological and lexical characters. In a series of tests on data from the Indo-European language family, they show that a perfect phylogeny is not possible. While most characters were compatible with a single evolutionary tree, there were inconsistencies localised particularly in the Germanic languages, presumably reflecting borrowing. Extensions of this method, reported in Nakhleh, Ringe, and Warnow (2005); Warnow et al. (2006), adapt the perfect phylogenies method to search instead for perfect phylogenetic networks. The perfect phylogenetic network infers a tree with a certain number of contact/reticulate branches which allow horizontal transfer events. The method seeks to maximise the amount of treelike character transmission (as per the previous perfect phylogenetic tree approach), and minimise the contact edges and transfer events.

In a complete departure from tree and network models, the STRUCTURE method (Pritchard, Stephens, and Donnelly 2000; Rosenberg et al. 2002) infers ancestral population structure by modelling evolutionary admixture. This method has been used by Reesink, Singer, and Dunn (2009) to investigate language dispersal and contact in the Sahul region, and explore the possible traces of ancient Australian-Papuan connections.

## 5 Testing hypotheses about language change

Quantitative phylogenetic approaches to language change should not be treated solely as an alternative to traditional methods in historical linguistics. As a tool for inferring language phylogeny, quantitative methods are inseparable from the Comparative Method. They are methodologically inseparable, because they rely on the same fundamental notions such as "cognacy", but also conceptually inseparable, because the Comparative Method provides a kind of gold standard for evaluating new phylogenetic results. But from another perspective, quantitative phylogenetic methods have much broader ambitions: that they are the first step for incorporating statistically framed models of historical relatedness into tests of hypotheses about evolved phenomena. In a rather unfortunate collision of terminology, these methods are called "comparative methods" in evolutionary biology (less often "The Comparative Method", but see Harvey and Pagel 1991). Phylogenetic comparative methods use phylogenetic trees to model other evolved aspects of language. This can involve inference of ancestral states, tests of dispersal order of languages, rates of change and evolutionary regime, coevolution of aspects of language, or of language and culture (Mace and Pagel 1994; **XREF:Greenhill**).

### 5.1 Dating and Phylogeography

The techniques for dating phylogenetic trees are closely integrated into Bayesian phylogenetic inference, as implemented by the BEAST package, although it is possible to log trees calibrated by substitution rates only. It is only a small conceptual step from substitution rates to chronological rates: if some tree-internal chronological calibration points are known (such as ancient inscriptions recording the arrival of a new ethnolinguistic group, or archaeological evidence dating the first arrival on an island), normalizing factors can be calculated for each branch according to what is required to morph the uncalibrated tree so as to match the calibration points.

The dates of internal nodes are specified as a probability distribution, and different types of distribution can express different kinds of prior knowledge. A lognormal prior can express a situation where we know that node $N_1$ must be dated before time $T_1$, but unlikely to be a long time before; a normal prior expresses knowledge that node $N_2$ must be around time $T_2$ with a certain standard deviation; a uniform prior expresses that node $N_3$ must be some time between $T_3$ and $T_4$. During the MCMC parameter search it is possible to log the inferred dates for the root and any other unobserved internal nodes of interest.

Other kinds of historical prior are also possible. In the phylogeographic approach a model of spatial diffusion is incorporated into the model, along with spatial priors (Walker and Ribeiro 2011; Bouckaert et al. 2012). This allows probabilistic inference of the spatial location of particular nodes in the past. As the spatial inference is incorporated into the likelihood calculation, this represents a principled violation of the principle of "Only Linguistic Evidence" (Campbell 1998: 323).

### 5.2 Character evolution

One of the principal uses of phylogenetic comparative methods is to model the evolution of characters along trees. The tree is generated independently to this analysis; in effect, the Bayesian tree sample is a phylogenetic prior in the analysis of the evolutionary behaviour of another character. Phylogenetic comparative methods can be used to infer ancestral states of a feature which is known to evolve along with the family. There are three requirements: (i) a genealogical hypothesis, (ii) a model of transitions between states of a feature (e.g. probabilities for transition from 'presence' to 'absence', and from 'absence' to 'presence'), and (iii) a set of observations of feature states from the tips of the tree. From these the method allows us to infer the most likely parameters of the model, and thus to make probabilistic judgements about the state of the feature at unobserved nodes of the tree. This is a standard technique used in computational phylogenetics, and has been applied in cultural evolutionary studies of kinship and residence systems in Austronesian (Jordan et al. 2009; Jordan 2011) and Indo-European societies (Fortunato, Holden, and Mace 2006; Fortunato 2011a, 2011b, 2011c), as well as to infer ancestral subsistence mode in Bantu-speaking societies (Mace and Holden 2005). Ancestral state reconstruction methods have barely begun to be exploited for purely linguistic questions.

Phylogenetic comparative methods can also be used to test for evolutionary dependencies; i.e. whether changes in one trait on a tree regularly correspond to changes in another trait. This technique (currently available as the DISCRETE test in the BayesTraits package) has been used to test evolutionary hypotheses in anthropology and linguistics. Holden and Mace (1997; 2003) use the method to show how the cultural innovation of pastoralism led to the evolution of lactose tolerance and patrilineal descent in Bantu societies (see also Mace and Holden 2005). Dunn, Greenhill, et al. (2011) use the same methods to test some of Dryer's generalizations about the Greenbergian word order universals.

### 5.3 Evolutionary ordering and evolutionary mode

I date the start of the modern phylogenetic era in linguistics to Gray and Jordan (2000), whose clever treatment of dispersal order as an ordered set of multistate features allowed them to

use a parsimony analysis to test the Taiwanese Origin hypothesis for Austronesian. Later, Bayesian analyses addressing the same question have improved upon these beginnings statistically, but have not produced a substantially different result (Gray, Drummond, and Greenhill 2009). Holden (2002) addresses a similar question for Bantu-speaking cultures, but in terms of subsistence rather than spatial location.

The best known study is Atkinson et al. (2008), which tested whether the Austronesian languages showed *punctuated evolution*, or jumps in the rate of change correlating with nodes on the tree. Punctuated evolution can be measured with *Pagel's* $\kappa$, one of Pagel's three comparative method statistics (Pagel 1997, 1999). *Pagel's* $\lambda$, along with *Blomberg's K*, tests for phylogenetic signal: the extent to which a trait (e.g. a feature of language) follows a pattern which is determined by phylogeny (Pagel 1997, 1999; Blomberg, Garland, and Ives 2003). Pagel's other phylogenetic comparative method test, *Pagel's* $\delta$, which tests whether rates of change increase or decrease over time, does not have any obvious application to any linguistic questions.

## 6 Conclusion

Recent advances in quantitative approaches to language history have put diachronic approaches at the forefront of modern linguistics. The power of computational phylogenetics and network analyses to perform explicit testing of evolutionary scenarios is only beginning to be explored.

## 7 Technical advice

MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) is a good, general-purpose Bayesian Phylogenetic Inference package, and is well-documented and reasonably easy to use. BEAST (Drummond et al. 2012) is more complex, but is also closer to the forefront of methodological development in phylogenetics, and the developers are active participants in language evolution projects. The Stochastic Dollo model in BEAST was implemented specifically for modelling the evolution of language. One of the difficulties in using BEAST is that it doesn't offer any default setting: the developers do not want to encourage researchers to use software without making informed decisions about priors. This uncompromising attitude is offset, however, by the friendly and helpful user community.

The phylogenetic packages provided for the R statistical programming language are highly flexible and the number of different packages is growing rapidly. The Phylogenetics Task View on CRAN (`http://cran.r-project.org/web/views/Phylogenetics.html`) is a continuously updated list of the different methods available.

## 8 Outstanding problems

Phylogenetic linguistics is an area undergoing rapid development, and we can look forward to improvements and further developments to many of the ideas described above. With respect to phylogenetic inference proper, we can expect to see more work on computational replication of the Comparative Method (detecting sound changes, inferring cognates and sound changes simultaneously). At the time of writing Bouchard-Côté et al. (2013) doesn't yet offer a usable methodology for non-developers to apply their techniques, but we can expect that this approach will gain in power and credibility. Borrowing and reticulation in trees has been addressed

(Nakhleh, Ringe, and Warnow 2005; Shijulal et al. 2011), but these methods are not currently mature enough to be used more widely: no standard software incorporates these methods, and they are not yet being used outside the teams that developed them.

Once phylogenetic network methods become more widely usable, then we can hope that it will become possible to begin using networks in phylogenetic comparative method hypothesis testing. Phylogenetic comparative methods are also developing rapidly at the moment, and the possibilities for using them to examine new linguistic questions in tree-like phylogenies are certainly not exhausted either.

Finally, we can always use better databases. This should include more, higher quality language data, as well as linguistic databases linked to databases from the cultural, geographical, ecological, historical, and biological domains. The Austronesian Basic Vocubulary Database (`http://language.psy.auckland.ac.nz/austronesian/`) and the Indo-European Lexical Cognate database (`http://ielex.mpi.nl/`) stand out here as large, accessible databases undergoing active development.

## 9 Further reading

- Gray, Greenhill, and Ross (2007) is a short survey of phylogenetic comparative methods and how they can shed light on linguistic questions.
- Nunn (2011) presents an accessible, book-length survey of phylogenetic comparative methods.
- Nichols and Warnow (2008) survey some of the quantitative methods used to infer linguistic phylogenies and test their performance against some sample language data.
- Lemey, Salemi, and Vandamme (2009) is a collected volume, with chapters treating all the major themes in phylogenetic inference.
- Paradis (2012) is a useful manual to help researchers implement their own phylognetic analyses using the R statistical programming environment.

## 10 Related topics

- Greenhill's chapter on the *Demographic correlates of language change* also discusses the problem of phylogenetic non-independence of observations of cultural and linguistic diversity.
- Chapters by Heggarty (*Prehistory through language and archaeology*) and Pakendorf (*Historical linguistics and molecular anthropology*) are concerned with linking historical linguistics to events and individuals existing in the past, through quantitative and statistical methods.

## References

Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill, and Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319 (5863): 588–588.

Blomberg, Simon P., Theodore Garland Jr., and Anthony R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57 (4): 717–745.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences.*

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337 (6097): 957–960.

Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences* 279 (1747): 4590–4595.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Vellupillai. 2008. Automated classification of the world's languages: description of the method and preliminary results. *STUF – Language Typology and Universals* 61 (4): 285–308.

Campbell, Lyle. 1998. *Historical linguistics: an introduction.* Edinburgh: Edinburgh University Press.

Daniëlsen, Swintha, Michael Dunn, and Pieter C. Muysken. 2011. The spread of the Arawakan languages: a view from structural phylogenetics. In *Ethnicity in ancient amazonia: reconstructing past identities from archaeology, linguistics, and ethnohistory,* ed. Alf Hornborg and Jonathan D. Hill, 173–196. Boulder: University of Colorado Press.

Donohue, Mark, and Simon Musgrave. 2007. Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 46 (2): 348–387.

Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4 (5): e88.

Drummond, Alexei J., and Marc A. Suchard. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology* 8 (1): 114.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29 (8): 1969–1973.

Dunn, Michael. 2009. Contact and phylogeny in Island Melanesia. *Lingua* 119 (11): 1664–1678.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson, and Neele Becker. 2011. Aslian linguistic prehistory: a case study in computational phylogenetics. *Diachronica* 28 (3): 291–323.

Dunn, Michael, Robert Foley, Stephen C. Levinson, Ger Reesink, and Angela Terrill. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46 (2): 388–403.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473:79–82.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. 2008. Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia. *Language* 84 (4): 710–759.

Dunn, Michael, and Angela Terrill. 2012. Assessing the lexical evidence for a Central Solomons Papuan family using the Oswalt Monte Carlo test. *Diachronica* 29 (1): 1–27.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309 (5743): 2072–2075.

Felsenstein, Joseph. 2004. *Inferring phylogenies*. Sinauer Associates Sunderland.

Fortunato, Laura. 2011a. Reconstructing the history of marriage and residence strategies in Indo-European—Speaking societies. *Human Biology* 83 (1): 129–135.

---------. 2011b. Reconstructing the history of marriage strategies in Indo-European—Speaking societies: monogamy and polygyny. *Human Biology* 83 (1): 87–105.

---------. 2011c. Reconstructing the history of residence strategies in Indo-European—Speaking societies: neo-, uxori-, and virilocality. *Human Biology* 83 (1): 107–128.

Fortunato, L., C. Holden, and R. Mace. 2006. From bridewealth to dowry? *Human Nature* 17 (4): 355–376.

Goebl, Hans. 1993. Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data. In *Contributions to quantitative linguistics,* ed. R. Köhler and B Rieger, 277–315. Dordrecht: Kluwer.

Gray, R.D., and F.M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405 (6790): 1052–1055.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323 (5913): 479–483.

Gray, Russell D., Simon J. Greenhill, and Robert M. Ross. 2007. The pleasures and perils of darwinizing culture (with phylogenies). *Biological Theory* 2 (4): 360–375.

Greenhill, Simon J. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics* 37 (4): 689–698.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics Online* 4:271.

Harvey, P.H., and M.D. Pagel. 1991. *The comparative method in evolutionary biology.* Oxford: Oxford University Press.

Heeringa, Wilbert Jan. 2004. Measuring dialect pronunciation differences using levenshtein distance. PhD Thesis, Rijksuniversiteit Groningen.

Heeringa, Wilbert, and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13 (03): 375–400.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across sub-saharan africa: a maximum-parsimony analysis. *Proceedings of the Royal Society B: Biological Sciences* 269 (1493): 793–799.

Holden, Clare, and Ruth Mace. 1997. Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* 69 (5): 605–628.

---------. 2003. Spread of cattle led to the loss of matrilineal descent in africa: a coevolutionary analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270 (1532): 2425–2433.

Holland, Barbara R., Katharina T. Huber, Vincent Moulton, and Peter J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution* 21 (7): 1459–1461.

Holman, Eric W., et al. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52:841–875.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754–755.

Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan E. Healy, George Koki, Françoise R. Friedlaender, and Jonathan S. Friedlaender. 2008. Genetic and linguistic coevolution in northern Island Melanesia. *PLoS Genetics* 4 (10).

Hunley, Keith, et al. 2007. Inferring prehistory from genetic, linguistic, and geographic variation. In *Population genetics, linguistics, and culture history in the southwest Pacific,* ed. Jonathan S. Friedlaender, 141–155. Vol. 1.

Jordan, Fiona M. 2011. A phylogenetic analysis of the evolution of Austronesian sibling terminologies. *Human Biology* 83 (2): 297–321.

Jordan, Fiona M., Russell D. Gray, Simon J. Greenhill, and Ruth Mace. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B: Biological Sciences* 276 (1664): 1957–1964.

Kassian, Alexei, George Starostin, Anna Dybo, and Vasiliy Chernov. 2010. The Swadesh wordlist. an attempt at semantic specification. *Journal of Language Relationship* 4:46–89.

Kass, Robert E., and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90 (430): 773–795.

Lemey, Philippe, Marco Salemi, and Annemieke Vandamme, eds. 2009. *The phylogenetic handbook.* Cambridge: Cambridge University Press.

Mace, Ruth, and Clare J. Holden. 2005. A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution* 20 (3): 116–121.

Mace, Ruth, and Mark Pagel. 1994. The comparative method in anthropology [and comments and reply]. *Current Anthropology* 35 (5): 549–564.

Manni, Franz, Etienne Guérard, and Evelyne Heyer. 2004. Geographic patterns of (genetic, morphologic, linguistic) variation. *Human Biology* 76:173–190.

McMahon, A.M.S., and Robert McMahon. 2005. *Language classification by numbers.* Oxford University Press, USA.

Nakhleh, Luay. 2011. Evolutionary phylogenetic networks: models and issues. In *Problem solving handbook in computational biology and bioinformatics,* ed. Lenwood S. Heath and Naren Ramakrishnan, 125–158. Springer.

Nakhleh, Luay, Donald A. Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* 81 (2): 382–420.

Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an indo-european dataset. *Transactions of the Philological Society* 103 (2): 171–192.

Nettle, Daniel. 1999. Is the rate of linguistic change constant? *Lingua* 108 (2–3): 119–136.

Nicholls, Geoff K., and Russell D. Gray. 2008. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (3): 545–566.

Nichols, Johanna, and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2 (5): 760–820.

Nunn, C. L. 2011. *The comparative method in evolutionary anthropology and biology.* Chicago: University of Chicago Press.

Oswalt, Robert L. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3 (3): 117–129.

---------. 1998. A probabilistic evaluation of North Eurasiatic Nostratic. In *Nostratic: sifting the evidence,* ed. Joseph C. Salmons and Brian D. Joseph, 199–216. Amsterdam: John Benjamins.

Pagel, Mark. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26 (4): 331–348.

---------. 1999. Inferring the historical patterns of biological evolution. *Nature* 401 (6756): 877–884.

Paradis, Emmanuel. 2012. *Analysis of phylogenetics and evolution with R.* 2nd ed. Springer.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Peer, Yves Van de. 2009. Phylogenetic inference based on distance methods. In *The phylogenetic handbook,* ed. Philippe Lemey, Marco Salemi, and Annemieke Vandamme, 142–180. Cambridge: Cambridge University Press.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155 (2): 945–959.

Prokić, Jelena. 2010. Families and resemblances. PhD Thesis, University of Groningen.

Pybus, Oliver G. 2006. Model selection and the molecular clock. *PLoS Biology* 4 (5): e151.

Reesink, Ger, and Michael Dunn. 2012. Systematic typological comparison as a tool for investigating language history. National Foreign Language Resource Center.

Reesink, Ger, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7 (11): e1000241.

Ringe, Don, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100 (1): 59–129.

Ronquist, F., and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Rosenberg, Noah A., Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. 2002. Genetic structure of human populations. *Science* 298 (5602): 2381–2385.

Schmidt, Heiko A., and Arndt Haeseler von. 2009. Phylogenetic inference using maximum likelihood methods. In *The phylogenetic handbook,* ed. Philippe Lemey, Marco Salemi, and Annemieke Vandamme, 181–209. Cambridge: Cambridge University Press.

Shijulal, Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences* 278 (1713): 1794–1803.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96 (4): 452–463.

Walker, Robert S., and Lincoln A. Ribeiro. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B: Biological Sciences* 278 (1718): 2562–2567.

Warnow, Tandy, Steven N. Evans, Donald Ringe, and Luay Nakhleh. 2006. A stochastic model of language evolution that incorporates homoplasy and borrowing. In *Phylogenetic methods and the prehistory of languages,* ed. Peter Forster and Colin Renfrew, 75–90. Cambridge: McDonald Institute for Archaeological Research.

**Index**