VERTICAL AND HORIZONTAL TRANSMISSION IN LANGUAGE EVOLUTION

WILLIAM S-Y. WANG & JAMES W. MINETT

APPENDIX A — LANGUAGE MODEL ALGORITHM

In order to generate pseudo-random character state data, we have constructed a model for language change among two sub-groups of genetically related languages belonging to a single language family. The model is encoded as a Windows-executable program, extending the algorithm of Minett & Wang (2003) for generating such data for sets of three languages. The algorithm accepts several parameters: the number of languages in each sub-group of the family, the time depth of the proto-language of the entire family, the time depth of the proto-language of each sub-group, the number of characters for which character states are generated, as well as parameters that describe the vertical transmission and horizontal transmission of the characters.

The key assumptions in the language model are as follows:

1. Topology: Languages bifurcate at a constant rate into two distinct derivative languages, each one, initially, having character states identical to the parent language. The derivative languages then evolve independently.

2. Vertical transmission: The *retention rate* of each character in each language is treated as a random variable, assumed here to be uniformly distributed. For example, one might specify the retention rate to lie within the range [80%, 90%]. Each character in each language is assigned a retention rate independently within that range, thereby

allowing heterogeneous retention rate to be modelled.[1] The retention rate may vary either across characters or across lineages, or across both.

3. Horizontal transmission: Modelled in much the same way as vertical transmission but in terms of a *contact rate* — one specifies a range for the probability that each character is acquired when a pair of languages come into contact. Multiple instances of contact can be injected at arbitrary time depths. In the experiments that we describe in the main text of this paper, only a single instance of contact has been injected except as explicitly noted. Furthermore, all instances of contact have been injected at zero time depth. At each instance of contact, all characters have a single opportunity to be acquired from the donor language by the recipient language.

The steps by which we model the language change are as follows:

The first step is to generate the genetic relationships among the specified number of languages. To do so, we assume that the two sub-groups of languages derive from a common proto-language. We "grow" a rooted binary tree, beginning with the proto-language located at its root. We allow the branches of the tree to bifurcate at a constant rate, each bifurcation adding an extra node, representing an extra language, to the tree. The first bifurcation forms the two sub-groups, which are then grown until they reach the specified sizes. Once the topology of the tree has been fixed, we re-scale the time depth associated with each branch so that the root has the specified time depth. Figure A1 summarizes this process.

---

[1] Empirically more realistic models of heterogeneous retention rate have yet to be incorporated into the model, such as treating the retention rate of each character as a gamma-distributed random variable (Cavalli-Sforza & Wang 1986; Gray & Atkinson 2003), by modeling the 'aging' of a character by decreasing its retention rate the longer it retains its state over time (Starostin 2000), or, in a similar way, by modeling 'cultural displacement' (Pagel 2000).

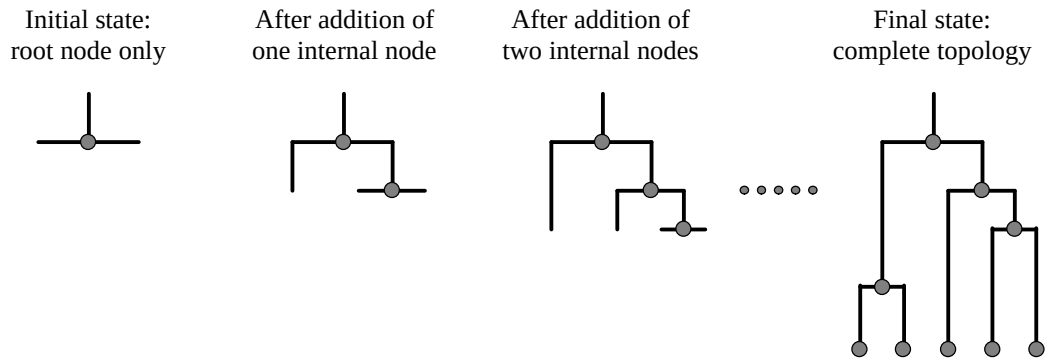

Figure A1. Topology “growth” process. The tree is grown from a single node (the root) one branch at a time. Branches bifurcate at a constant rate, each bifurcation adding one additional node (adding an extra language) until the tree has the required number of terminal nodes (languages). Once the topology has been fixed, the branch lengths are re-scaled so that the root has the specified time depth.

The second step is to generate the character states of each language. The proto-language (root node) has the state zero assigned to each character. The tree is then traversed node-by-node (by pre-order traversal) allowing the state of each character to be replaced by a new, unique character state with probability according to the requested probability of retention and the time depth of the parent branch — we use the standard glottochronological formula for the probability of character retention, $p = r^t$, where p is the probability of retention, t is the time depth and r is the probability of retention per unit of time. The retention rate is treated as a random variable with uniform distribution on some specified range, e.g. [80%, 95%]; each language is independently assigned a separate value of the retention rate to each character. Characters that are replaced are assigned a new, unique state. Figure A2 summarizes the process of character state allocation.

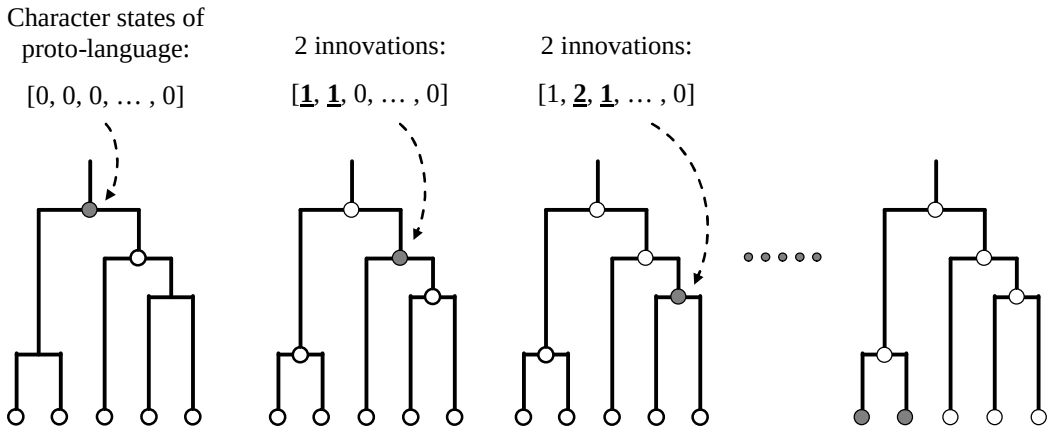

Figure A2. Character state allocation process. The characters of the proto-language (root node) are all initialised to the state 0, represented in the figure by the vector [0, 0, 0, … , 0]. The tree is then traversed one node at a time, updating the character states according to the probability of retention and the relative time depth of the parent node. Each time a character undergoes replacement, a unique state is assigned to it. Character state changes are represented in the figure by underlined boldface.

The third step is to model the language contact. For each requested contact event, one donor and one recipient language is selected, one language taken from each of the two sub-groups. The contact rate, which is treated as a random variable in much the same way as the retention rate, specifies the probability that the donor state of each character is adopted by the recipient. Each character is assigned a probability independently. Six contact scenarios involving up to two contact events are modelled, as described in Section 3 of the main text.

The input parameters of the algorithm are:

- the number of languages in each sub-group of the family;
- the time depth of the proto-language of the entire family;
- the time depth of the proto-language of each sub-group;

- the number of characters;
- the mean retention rate of characters;
- an interval describing the heterogeneity of retention rate across characters;
- an interval describing the heterogeneity of retention rate across lineages;
- an interval describing the probability that a character is acquired in one instance of contact;
- the number of instances of language contact. In each of the experiments reported here, all contact events occur at zero time depth, modelling recent instances of language contact.

The algorithm requires that precise values be specified for each parameter value. However, when running the algorithm to test the performance of the skewing method for particular sets of languages, few of these parameter values are available to the linguist. How then should the linguist proceed? The approach that we suggest is to specify upper and lower bounded estimates for the unknown parameters, particularly the time depths of the proto-language of the entire family and of each sub-group. The performance may then be tested for different combinations of these parameter values to estimate a lower bound on the performance; the narrower the bounds on the estimated parameter values, the greater the performance.

REFERENCES

CAVALLI-SFORZA, LUIGI LUCA & WANG, WILLIAM S-Y., 1986. 'Spatial distance and lexical replacement', *Language* 62, 38–55

GRAY, RUSSELL D. & ATKINSON, QUENTIN D., 2003. 'Language-tree divergence times support the Anatolian theory of Indo-European origin', *Nature* 426, 435–439.

MINETT, JAMES W & WANG, WILLIAM S-Y., 2003. 'On detecting borrowing: distance-based and character-based approaches', *Diachronica* 20:2, 289–330.

PAGEL, MARK, 2000. 'Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Vol. 1, Cambridge: The McDonald Institute for Archaeological Research, 189–207.

STAROSTIN, SERGEI, 2000. 'Comparative-historical linguistics and lexicostatistics', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Vol. 1, Cambridge: The McDonald Institute for Archaeological Research, 223–259 (translation from Russian by N. Evans and I. Peiros).