

Target authors' response

Universal typological dependencies should be detectable in the history of language families

STEPHEN C. LEVINSON, SIMON J. GREENHILL, RUSSELL D. GRAY and
MICHAEL DUNN

1. Introduction

We claim that making sense of the typological diversity of languages demands a historical/evolutionary approach. We are pleased that the target paper (Dunn et al. 2011a) has served to bring discussion of this claim into prominence, and are grateful that leading typologists have taken the time to respond (commentaries denoted by boldface). It is unfortunate though that a number of the commentaries in this issue of *LT* show significant misunderstandings of our paper. **Donohue** thinks we were out to show the stability of typological features, but that was not our target at all (although related methods can be used to do that: see, e.g., Greenhill et al. 2010a, Dediu 2011a). **Plank** seems to think we were arguing against universals of any type, but our target was in fact just the implicational universals of word order that have been the bread and butter of typology. He also seems to think we ignore diachrony, whereas in fact the method introduces diachrony centrally into typological reasoning, thereby potentially revolutionising typology (see **Cysouw's** commentary). **Levy & Daumé** think we were testing for lineage-specificity, whereas that was in fact an outcome (the main finding) of our testing for correlated evolution. **Dryer** thinks we must account for the distribution of language types around the world, but that was not our aim: our aim was to test the CAUSAL CONNECTION between linguistic variables by taking the perspective of language evolution (diversification and change). **Longobardi & Roberts** seem to think we set out to extract family trees from syntactic features, but our goal was in fact to use trees based on lexical cognates and hang reconstructed syntactic states on each node of these trees, thereby reconstructing the processes of language change. Many commentators think that the methods are flawed in principle – they use the wrong

trees (**Donohue, Plank**), fail to take account of contact (**Donohue, Bickel**), amplify noise (**Baker**), and so on. While the methods are relatively new in linguistics and can no doubt be improved, we don't think they have the fatal flaws attributed to them. It is perhaps unfortunate that the rapid development of evolutionary methods in linguistics has taken place largely in a parallel literature, but nevertheless there are substantial introductions and explications for linguists (see, e.g., McMahon & McMahon 2005, Dunn et al. 2008, Dunn 2009, Greenhill & Gray 2009, Gray et al. 2011).

Perhaps in part these particular misunderstandings can be put down to the brief, technical nature of the paper. But in part they also seem to reflect a mistrust of hi-tech methods in typology (see, e.g., **Baker, Donohue**) which is, we think, misplaced. What we were in fact trying to do is find a way of sifting Greenbergian universals so that we can decide which of them are "accidental" correlations, and which of them seem causally linked, so that they could actually constrain language history, presumably by channelling the cognition that enables transmission. The filter we employed was diachronic: genuinely linked variables should show causal binding across time, during language change and diversification.

In the following section we try to explain the basic underlying reasoning, turning in the remaining sections to some of these recurrent points of contention. In the final section, we collect some responses to points made in individual commentaries that did not fit neatly into the body of our response.

2. Two strategies for finding linked traits, implicational universals, or parametric variation

A tremendous amount of collective energy has gone into building the knowledge we now have about the distribution of linguistic diversity – the project of language typology. The investment has made it possible to ask fundamental questions about the nature of linguistic diversity, and is one of the triumphs of modern linguistics. The patterns revealed are the product of many factors, but (as **Dryer** agrees) very largely result from historical processes of diversification and diffusion across the geography of the inhabited world.

There are many different projects that can be pursued with this data, as illustrated by the two target papers in this issue, which have quite different goals. Atkinson's (2011) paper, also a target for commentaries in this issue, explores possible traces of the original diaspora of modern humans. Our paper examines whether there are fundamental constraints on language change of the kind presumed in mainstream theoretical linguistics – constraints that would presumably be cognitive in nature and operative through the processes of transmission across generations. Atkinson's paper has a historical goal, while ours

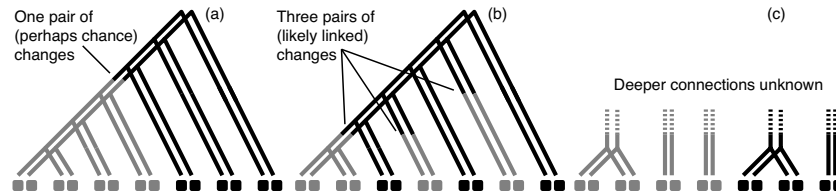


Figure 1. Correlated pairs of features indicated by the pairs of shaded squares. The histories creating this distribution are shown by the doubled trees. The different histories lead to very different evaluations of how likely these correlations are to be causal. In (a) the correlation is the product of a pair of changes on a single branch, which provides much weaker evidence for a causal, non-chance, linkage than the three pairs of changes which produced the distribution of features states shown in (b). Where the ancestral connections between languages are unknown (c) the grounds for concluding causality are weaker still.

uses history to get at the presumed cognitive constraints thought to govern the possibility space that languages may occupy.

Unlike Atkinson's paper, our paper thus goes for the jugular of linguistic typology. Mapping the diversity of languages has many uses, but the main motivation for the development of typology was surely the search for constraints on possible human languages – for systematic pressures that channel diversity into certain pathways. The search for these constraints, for invariants and biases, has gone under the rubric of “Universals”, and has used sophisticated methods for sampling the languages of the world to extract the invariants or the major trends. In practice it has focused on implicational universals, and every textbook highlights the word order implicational universals – the target of our article – that have been tested on a worldwide sample.

There are quite significant difficulties for uncovering universals, the most pertinent being “Galton's problem” (see, e.g., Mace & Pagel 1994). This is the difficulty of distinguishing between two causally-linked traits versus two traits that simply share a common ancestor (or have both been borrowed from an extinct common neighbour; see again **Cysouw's** commentary). Features that are shared between languages merely due to historical relationships cannot be treated as independent pieces of evidence. Consider the histories of the pairs of features mapped in Figure 1. In both (a) and (b) there is an apparently perfect correlation of the two features. But the history behind the distribution of states in (a) reveals that the correlation is the product of a single pair of changes. The distribution of states in (b) is the product of three pairs of correlated changes, and the probability that this reveals a causal mechanism linking these states is much higher.

Two specific issues relevant to Galton's problem arise in the linguistic domain: (i) a few large language families account for most of the languages in the world (the four families in the study exhaust over a third of the world's diversity), and (ii) there is every reason to think that ultimately all extant languages descend from one or only a few ancestral sources. A typical response to Galton's problem is a stratified-sampling-based approach like **Dryer's**: (i) include related languages, but take only one from each "genus" or major branch of a family, (ii) look for patterns shared across most continental landmasses (with an implicit pre-1492 baseline). The first tactic obviously falls afoul of Galton's problem, since it treats related languages as independent samples. The second hardly controls for shared inheritance either, since we know that, for example, the Americas were relatively recently colonized at a time depth not so much greater than, for example, Indo-European (witness the recent apparent discovery of cross-continental connections between Yeniseian and Na Dene; Kari & Potter 2010). **Dryer's** areas also wobble between geography and known familial connections – many of the languages of New Guinea are of course Austronesian, thus Asian in origin, so as partial remedy South-East Asia is lumped with Oceania (see **Bickel's** commentary which uses different areas). Overall, this does not look like a firm foundation for the inference of causal connections between linked traits, as **Cysouw** points out.

Many typologists may assume that the dangers of covert phylogenetic dependence are remote. But given the apparent genetic bottlenecks at the beginning of the modern human diaspora out of Africa (Amos & Hoffman 2009), something close to language monogenesis seems a reasonable assumption, rendering Galton's problem insurmountable. Furthermore, typological features can be remarkably stable over the life of whole language families (witness the case of Bantu in our study), so typological settings are likely in part to outdate known families (see Dunn et al. 2005, Reesink et al. 2009, Greenhill et al. 2010a, Dediu 2011a). We therefore think there is every reason to take covert phylogenetic dependence seriously.

There is an alternative strategy: to properly control for phylogenetic relatedness, and to look for evidence of causal connections between two traits. A causal connection between two parameters predicts that when the value of one changes, so does the value of the other: Specifically, if we look *WITHIN* a language family and trace the history of traits, and find that whenever the value of parameter A changes so does the value of parameter B, then we have a statistical argument for a causal linkage between the two parameters. The more independent instances of coupled changes, the stronger the statistical support for the causal connection. Quantifying the probabilities of causal linkage is quite straightforward here compared to the task of estimating chance co-occurrences in a worldwide hand-picked sample with unknown genealogical connections. The commentary by **Cysouw** nicely explains the logic of this approach.

Our study followed the strategy of testing whether the Greenbergian word order universals, whatever apparent confirmation they get from worldwide correlations, also constrain language change within families. If the Greenbergian generalisations manifest themselves also as causal linkages in language evolution governing diversification and change, then they would indeed constrain the space of possible languages, and presumably must be a product of the cognitive processes involved in language transmission. If not, then the Greenbergian generalisations may be no more than statements about the skewed historical distribution of extant languages.

The question we asked then was: Are Greenbergian implicational universals mere summaries of observed correlations, based on any number of historical factors, or can we find evidence for true causal linkage between variables across time, so that when one feature changes the correlated feature changes, revealing genuine causal linkage? Our results show that the word order correlations do not stand up to this more stringent test. Therefore, we reasoned that such correlations are not likely to operate as cognitive biases influencing language transmission. Since many linguists have supposed that such implicational universals reveal something important about the nature of the human language capacity, our findings seem to have fundamental importance for the field of typology.

Many of the commentaries seem to not fully grasp the complete change of strategy involved in this way of testing typological generalisations for underlying causal linkages:

- (i) Our study is necessarily diachronic, generalising over changes in the historical development of language families.
- (ii) It is necessarily tied to specific language families and their reconstructed history.
- (iii) It looks for causal connections between traits. Specifically, it asks the question whether one word order (between one pair of phrases) is linked to another word order (between another pair) in such a way that the change in one order triggers a change in the other.
- (iv) The study presumes that if such causal connections could be found, they would most likely have a cognitive basis that biases transmission, and play an important role in explaining distributional patterns across the languages of the world.
- (v) But if they cannot be found, those distributional patterns would still need explanation, presumably in purely historical terms.

We think this strategy has much to recommend it over the traditional worldwide sample:

- (i) We avoid the very real sampling difficulties of the worldwide sampling method.
- (ii) We directly examine the diachronic changes presumed implicitly in the Greenbergian method (see **Croft et al.** commentary).

- (iii) When we find dependencies, we can make detailed inferences about the directionality and timing of the changes.

3. Methods

There were a number of serious misunderstandings about what our methodology involved. We will take the opportunity here to outline in a little detail the three major steps of the analysis described in the target paper. Additionally, readers are referred to the supplementary materials to Dunn et al. 2011a which provide a basic description of the model testing approach, and Dunn et al. 2008, Dunn 2009, Greenhill & Gray 2009, Dunn et al. 2011b for introductions to Bayesian phylogenetic methods applied to language. (Outside linguistics these methods are very well established; see Huelsenbeck et al. 2001, Felsenstein 2004: Chapter 18 for useful expositions.)

The technique used in our paper detects evolutionary dependencies between features of languages in a way that controls for Galton's problem (that apparent correlations between features may be an artefact of genealogical structure) by looking at languages known to be related in specific ways.¹ There are three elements to our analysis: (i) a genealogical control (in this case, a set of phylogenetic trees), representing the known relationships between the languages; (ii) a set of observations of the structural states of extant or recorded languages (in this case, constituent-order features); and (iii) a statistical, probabilistic test of competing models of how these features have most likely evolved given the phylogenetic history. We will take these three steps in turn.

3.1. *Trees*

Linguists have tended to represent their knowledge about language family structure in terms of a single tree. However, such a single tree cannot do justice to the inevitable uncertainties that arise in making historical inferences with finite data (e.g., Felsenstein 1988, Huelsenbeck et al. 2000). In contrast, the genealogical control used in our analyses is, for each language family, a SET of trees made from basic vocabulary data. The set of trees enables us to quantify the uncertainty in the language subgrouping in a way that is not possible in a single tree. The basic vocabulary data used in our analyses consist of wordlists that have been cognate-coded by historical linguists. This means that they are true cognates, not mere "lookalikes". It also means that obvious borrowings have been removed from the analysis. We code these cognate sets in a binary matrix reflecting the presence or absence of a cognate in a language. We then build phylogenetic trees from this matrix using a model-based, probabilistic

1. Confusingly for linguists, this is known to other disciplines (biology, anthropology) as "the comparative method" (Harvey & Pagel 1991).

method called Bayesian Monte Carlo Markov Chain phylogenetic inference. The algorithm simultaneously searches for the most likely tree structure along with the most likely parameters of lexical change (representing the probability that a cognate set comes into being, and the probability that a reflex of a cognate set be lost). This is computationally intensive:² the Bayesian Monte Carlo Markov chain (or MCMC) part of the name stands for a family of techniques which make it feasible to estimate these values by making a non-exhaustive search of the space of possible parameters/topologies, taking samples in proportion to their probability given the model and the data. Under this technique, “good” explanations of the data recur many times in the sample, “bad” explanations are rare. The output from the MCMC search is thus a set of trees sampled in proportion to their posterior probability.

The tree sample gives us a measure of confidence in different aspects of the trees: some linguistic subgroups are present in every tree of the posterior sample, so we can be very confident in their reconstruction. There might be conflicting evidence for other elements of the reconstruction, and these competing hypotheses are represented in the sample in lesser proportions. The information contained in the tree sample is hard to visualise: in our original paper we used a *MAXIMUM CLADE CREDIBILITY TREE*, a single tree which is representative of the entire tree sample (selected by comparing the aggregate probability of all the clades in each tree of the sample). We chose this representation as it is common in phylogenetic studies, and is easy to interpret: in a maximum clade credibility tree each branch is annotated with a posterior probability value that denotes the strength of the evidence in the data for that node. So, a branch with a posterior probability of 1.0 is present in 100 % of the trees in the sample, whilst a branch with a posterior probability of 0.5 is only found in half. Another strategy for representing a tree sample with a single tree is the *CONSENSUS TREE*, a tree constructed from the most frequently occurring clades in the entire sample but that, unlike the maximum clade credibility tree, might not have occurred in the tree sample itself. However, we can represent the posterior tree sample in a different way to make the set of trees more vivid using a *DENSITREE* (Bouckaert 2010), a tree sample representation method that superimposes all the trees in the sample into a single figure (Figure 2).

Model-based tree inference produces tree topologies with branch lengths that express the amount of evolutionary change, which is the product of the

2. The number of possible trees for even a small number of related languages is quite surprising. For example, there are 213 billion possible bifurcating trees for a phylogeny with just 15 languages or species (Felsenstein 1978).

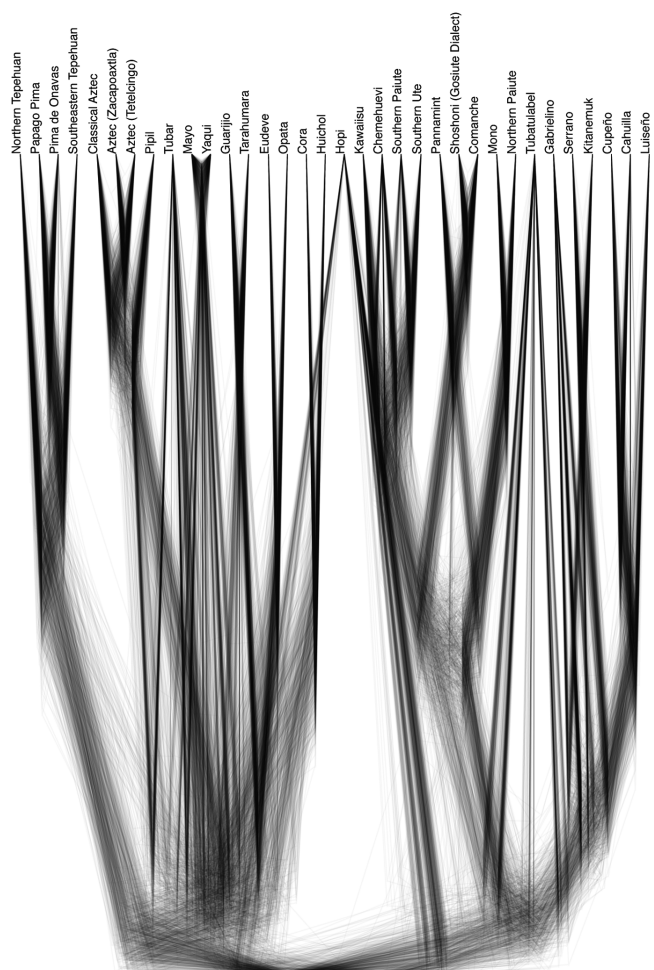


Figure 2. A DensiTree (Bouckaert 2010) representation of the 800 trees in the Uto-Aztecan tree sample: all the trees in the sample overlaid in transparency, so that areas of greater agreement show darker (for clarity, variation in branch lengths are not shown in this figure).

rate of change and the length of time represented by the branch.³ The MCMC parameter search estimates optimal values for both these parameters at once.

The kind of methods used in our study require family trees that have the following properties:

- (i) they are not based on syntactic features, since we need to map earlier states of syntactic parameters onto independently derived trees to avoid circularity (*pace* **Longobardi & Roberts**);
- (ii) they precisely quantify uncertainty in the nodes and branches and overall topology of the trees;
- (iii) they quantify the amount of evolution in each branch of the tree, allowing different rates of evolution across the tree.⁴

The kinds of trees found in studies using the comparative method, or in many standard linguistic references, are not quantified in this way, and do not reflect uncertainties and conflicting signals. Thus, despite the scholarship that goes into them, they remain in this respect fundamentally misleading.

The Bayesian phylogenetic analysis of cognate sets from basic vocabulary to construct trees overcomes these limitations and is (*pace* **Donohue**) a well-established method in its own right (Gray & Jordan 2000, McMahon & McMahon 2005, Gray et al. 2009, Greenhill & Gray 2009, Gray et al. 2011). As in most such studies, we used lexical data from Swadesh lists, a well-documented, conservative, part of the vocabulary with reasonably low rates of borrowing (see Haspelmath & Tadmor 2010). There are measures of reliability under different amounts of contact and lexical diffusion (Greenhill et al. 2009, Currie et al. 2010). As **Croft et al.** point out, adding phonological and morphological traits is not likely to substantially alter the results. Moreover, the resulting posterior sample of trees closely matches those derived by the comparative method. For example, a Bayesian phylogenetic analysis of Austronesian basic vocabulary recovered 26 out of 34 standard subgroups showing high concordance with the results of the linguistic comparative method (Gray et al. 2009, Greenhill et al. 2010b).

There were a number of unfortunate misunderstandings even of this preliminary stage of our analysis:

Longobardi & Roberts completely mistook our enterprise: they thought we were using syntactic variables to make trees, whereas the trees play only a background, if crucial, role in the study. What we were testing, against the

3. It is possible to extract the estimates of rate of change parameter and draw a tree where the branch lengths show time directly, but this is a different enterprise (Drummond & Rambaut 2007).

4. **Croft et al.** mistakenly say we did not use variable rates: we used the covarion method in the construction of the trees, as described for the Austronesian trees in Gray et al. 2009. See Section 4.2 for discussion.

background of trees, was whether values on one word order parameter governed values on another over the course of the evolution of a language family.

Donohue expresses several elementary misunderstandings of the nature of phylogenetic trees generated by Bayesian phylogenetic methods. He complains about the “binary branching” of our trees claiming that it is a completely unrealistic model. What **Donohue** doesn’t grasp is that a graphical summary of the posterior sample of trees cannot do justice to what is actually computed over – namely a forest of trees (but see Figure 2 as an attempt). Uncertainty about the branching is directly represented in the underlying sample. Each tree in the sample represents a possible history, and the elements of structure shared by the trees in the sample gives a measure of phylogenetic (un)certainty. This means that we can be maximally confident of branches found in all trees of the sample, and proportionally less certain as the sample shows more variation. The length of the branches on these trees is also meaningful, indicating (in our trees) the amount of evolution on each branch. As with the bifurcations of the trees, the tree sample gives us a statistical sample of the range of likely lengths of each branch. While a trifurcation is not in principle possible using these methods, in practice a very short branch can indicate the same relationship: if the data supported a trifurcation, then this is what the algorithm would find.

The issue of the placement of ancient languages as sisters of their clade is similarly trivial. Together with **Plank**, **Donohue** complains that ancient languages are placed on branches descending from the root nodes of their clades, rather than on the root node itself; so, for, example Ancient Greek is represented as a sister of Modern Greek. But, of course, Modern Greek didn’t descend directly from the literary Attic Greek represented in most ancient sources, but from some common ancestor. Our trees show the history of languages as modelled by the birth of cognate sets and death of lexical reflexes, inferred from lists of lexical cognate judgments. The trees are representations of the evolutionary relationships between these concrete lists, rather than relationships between languages as idealised, abstract entities. The Italian wordlist does not descend from the Latin wordlist in the same way that we descend from a grandparent. Rather, the Latin wordlist is a “doculect” (to use Cysouw’s terminology) closely related to the node of the tree that the whole of Romance is descended from. **Donohue**’s claim that our trees bear little resemblance to known genealogical relations is simply mistaken.

Plank is also mistaken in thinking that our method does not distinguish between shared retentions and innovations. Modern phylogenetic methods grew out of the realisation by Hennig (1966) that it was critical to discriminate similarity due to shared innovations and similarity due to retention. The phylogenetic methods we used to construct the language trees infer, in a probabilistic fashion, cognate gains and losses on every single tree (e.g., Felsenstein 1981,

Steel & Penny 2000). An accessible discussion of this issue is available in Greenhill & Gray 2009.

3.2. *Typological data*

The second step in the process required for this study is to tabulate the observed states of the typological features of interest. Like many other authors we owe an enormous debt to the indefatigable Matthew Dryer. We extracted data about eight different word order features from Dryer's contributions to the *World atlas of language structures* (Dryer 2008a–h), and added more information from other published sources. The sources of additional coded information, the states themselves, and details about how we coded these features are all to be found in the supplementary materials to Dunn et al. 2011a. We recoded these in a deterministic manner from the Dryer/WALS coding.

Any large scale undertaking of this kind is likely to raise queries from language specialists. Two commentaries (**Longobardi & Roberts, Baker**) in particular worry about the “surfacey” character of the coding: Baker gives the example of demonstratives that are adjectives, co-occurring with articles, and thus might trivially be expected to adopt the main adjective-noun order. But this brings up another central dilemma of typology: every language's categories are, in detail, language-specific in character, and the typologist's craft is to find good enough cover categories that make it possible to compare languages (see discussion in Haspelmath 2010). We think that the fine coding issues should then come up in the analysis of results, where specialist knowledge will help us to unpack what “lineage-specificity” really is and where it comes from. The different transition probabilities in Indo-European and Austronesian, illustrated in Figure 3 of the target article, almost certainly have their explanation in precisely this kind of detail, for example, the different diachronic sources for adpositions in the two families.

3.3. *Model testing*

We have explained how the family trees are built from lexical data, and how typological data for all the living or recorded languages have been collated and coded. The final step of the process is to reconstruct how these observed structural features might have evolved within the genealogies we have evidence for. That is, we need to infer how these states have evolved over all the ancestral nodes of the trees, taking into account that each ancestral node and its placement in a tree is itself only a probability. Crucially, we want to know whether a model that has an intrinsic coupling between a pair of word order features represents a more likely evolutionary story than a model where there is no such linkage, and the pair of features evolve independently. A model of feature evolution consists of a mathematical expression of rates of change (transitions)

between the different states of the features. The process of estimating these evolutionary parameters is conceptually similar to the process of Bayesian phylogenetic inference described in Section 3.1, except that instead of estimating tree topology along with the other parameters, the algorithm integrates over the phylogenetic trees inferred from the lexical data, fitting the structural data onto those trees. The results are once again a set of possible solutions to an evolutionary outcome sampled according to their probability, but in this case the sampled values are the transition parameters most likely to produce the observed set of typological states given the genealogy. The overall likelihood of the different models of evolution (i.e., dependent and independent) are compared using Bayes Factors, a statistic that expresses how much one hypothesis is to be preferred over another (Kass & Raftery 1995).

Many of the better informed commentaries concentrate on this stage of the analysis. **Levy & Daumé** provide a few useful intuitive characterisations of Bayes Factors, which have no direct counterpart in the more familiar *p*-values. They point out correctly that lack of support for a model can't be directly equated with the absence of the phenomena modelled, for all the normal reasons of sample size and the special reasons of lack of sample diversity (as in the case of Bantu). What they don't draw attention to, however, is that our sample consists of hundreds of EVENTS, instances of word order change, where that size is a function not only of language numbers, but age of family and rates of change.⁵ **Croft et al.** describe some ways in which a power analysis might be done – that is, some assessment of the minimal sample size required to find an effect. These are useful suggestions. We think our sample size is large enough for any reasonably strong bias to shine through. If a cognitive bias underlying Greenbergian “harmony” turns out to be vanishingly weak, and only detectable with massive samples, that itself is surely deeply telling: harmonic principles then would play only a very minor role, if any, in restricting the space for possible human languages. That in fact is our conclusion.

To illustrate that even with a relatively small sample the method returns decisive findings, consider the case of the linkage we discovered between object-verb order and subject-verb order exclusively in the Uto-Aztecan family. The point can be made graphically, as in Figure 3, which represents the two word orders mapped onto the same tree (once again a warning: this is a maximum clade credibility tree represented twice, and as explained is just one way of summarising the agreement in the forest of likely trees). Visual inspection shows that when the word order on the left varies, so does the word order on the right,

5. A maximum parsimony calculation of the total number of character state changes in the maximum clade credibility trees for our four language families gives the following estimates: Austronesian 107 changes; Bantu 12 changes; Indo-European 83 changes; Uto-Aztecan 37 changes.

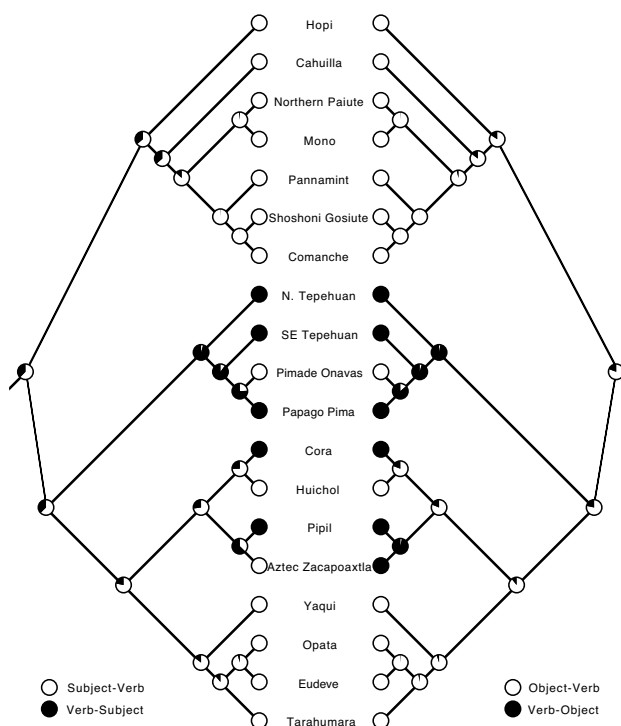


Figure 3. Maximum likelihood reconstructions of ancestral states of Subject-Verb (left) and Object-Verb (right) order features in the Uto-Aztecan languages estimates on the maximum clade credibility tree. Pie-charts on internal nodes show the probability of the inferred state at that node. For example, the probability of Proto-Uto-Aztecan having Subject-Verb is about 2/3. The visual impression that changes in the state of one feature tend to be coupled with changes in the state of the other is confirmed by the DISCRETE test reported in Dunn et al. 2011a (languages with missing or polymorphic features are omitted from this expository figure; in the real analysis polymorphic states are included as they contribute to the calculation of parameter values).

and in fact the computational method returns a high Bayes Factor that captures the corresponding likelihood of linked evolution.

Croft et al. wonder whether our linkage model (where two aspects of word order change together) unfairly lumps biconditional implications with one-way implications of the Greenbergian type. But the dependent model does not force the implications to be biconditional, and in fact far from obscuring one-way implications, our approach can actually be used to infer them, so long as these implications are interpreted diachronically. Thus, Figure 3 in the target article

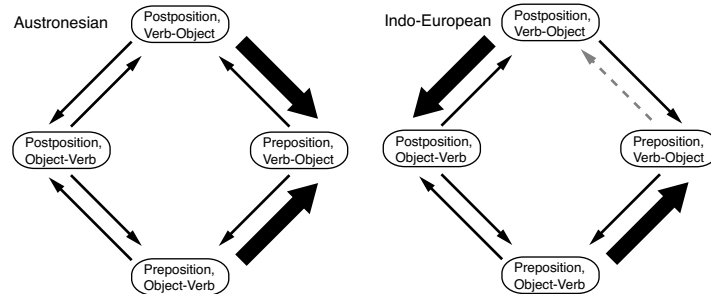


Figure 4. *Inferred transition probabilities for Adposition type and order of Object and Verb in Austronesian and Indo-European, reproduced from Dunn et al. 2011: 81. The weight of the arrow is proportional to the probability of the change (the dashed arrow indicates a probability estimated as zero).*

(reproduced here as Figure 4) shows that in Indo-European, verb-object order is much more likely to change to harmony with adposition type than vice versa (in Austronesian there is no such one way tendency, and only a high probability of entering and low probability of leaving the preposition, verb-object state).

We treated each of our language families as an independent experiment. A number of the comments search for a way to aggregate the results across the language families. **Levy & Daumé** constructively suggest a Bayesian method for testing whether pairs of language families are best described by a single or different models, i.e., for lineage-specificity. They also wonder, like **Croft et al.**, whether one couldn't just join all four families in one tree, on the assumption of deep monogenesis, and rerun all the analyses – if the independent model still seems better, then collectively there is no support for the linked word order (harmony) model. (Of course, as **Croft et al.** point out, lineage-specificity would now mean branch-specificity – but the interest would be the comparison between our original procedure and the new one.⁶) This suggestion seems interesting, perhaps, but it runs into serious objections. First, our trees are based on cognate sets, and of course there are none across the four families. Second, given that absence of information, how should one decide the branching: on geography, likely migration patterns, or what? Third, what kind of depth and rate would one imagine: back to the diaspora out of Africa at 70,000 years ago? And how would one correct for the lack of diversification except in the last 10,000

6. Croft et al. actually say that the BayesTraits software can't be used on monophyletic groups – we suppose this is just a slip of the pen, but it is important to correct. Monophyletic clades are simply groupings that contain all the descendants of a given node. Bayes Traits is perfectly happy with these.

years? This kind of speculation awaits a full-scale simulation of linguistic diversification (or an accepted classification of “Proto-World”). Meanwhile, the method we have used is a conventional, well-understood tool for testing phylogenetic dependency.

Croft et al. wonder whether our claim for lineage-specificity might seem to imply polygenesis – that is, more than one Ur-ancestor for all known human languages. We simply took it as a practical matter that our four language families, remote in space, were related, if at all, only in deep time. To repeat: due to the method we employed, constructing trees from lexical cognates, it is not possible to obtain any sensible structure by collapsing the four language families into one Ur-family. But for the record, we think monogenesis or at least “paucigenesis” a reasonable presumption, which is precisely why we think that Galton’s problem should be taken seriously by typologists.

Cysouw takes a close look at the statistical background for our methods. As we had clearly flagged, the methods used test for the support for correlated evolution (linked word order variables) – they can’t test for the strength of the contrary claim, that word order variables are independent of each other. He proposes a re-analysis of the data, using *recherché* meta-statistics that derive strength of support from a number of independent tests. This allows **Cysouw** to combine the results across the four lineages and see whether, in combination, there is support for the Greenbergian linkages between word orders. He claims that there is “clear agreement across lineages” although “these agreements are indeed not very strong”, requiring an extension of the sample beyond our four language families.

Two comments: First, as **Cysouw** points out, these Bayesian methods do not directly equate to standard statistical norms (for example, with respect to known degrees of freedom). So the correct statistical procedures for a meta-analysis are not straightforward, and are a matter for research in statistical theory. We chose to use the conventional interpretation of Bayes Factors (Kass & Raftery 1995). Given the weakness of any evidence for Greenbergian linkages, meta-analysis will inevitably suggest increasing the sample size, here understood as analysing more language families, and by that slippery slope we may tumble back into the problems of traditional typology and how to construct a balanced sample (see our remarks on **Bickel** below).

Second, reconsider the underlying hypothesis we are interested in: Is there a cognitive bias of some kind that channels languages into “harmonic” states? That is what we have taken to be the underlying motivation for much typological work. Now consider that we have examined hundreds of diachronic events in our four language families, and have not found striking evidence (as **Cysouw** agrees) for any such cognitive biases. As the signal for universal cognitive effects becomes weaker so their effect becomes increasingly over-ruled by local factors (see here **Dryer**’s admission that “Dunn et al. are right that cultural

evolution is more important than cognitive factors in governing word order change”). It is worth bearing in mind here the very cogent arguments based on experiment and modelling that if there is indeed a cognitive bias it is likely to be amplified greatly through cultural transmission (Kirby et al. 2007, Dediu 2011b). This makes **Cysouw**’s “not very strong” signal all the more surprising for Greenbergian theory. Perhaps we would do well to pause, reconsider what the goals of language typology really are, and what exactly we will take as evidence for the hypothesis we espouse.

Perhaps in this light the commentary by **Tily & Jaeger** will no longer seem to be a diversion from the central goals of typology. They point out that behavioural experiments, specifically varieties of artificial language learning, provide a very useful ancillary line of evidence for cognitive biases. Iterated learning experiments allow, just as in our method, the inspection of earlier states and how they are transformed. These experiments have their own Achilles’ heels of course, since every human already has a language, which will itself bias learning of a second one. Nor can this bias be removed simply by exploring artificial languages of a different pattern than the native one. We suggest that an interesting line of work would be to extend these experiments from “WEIRD” undergraduates (Henrich et al. 2010) to speakers of really different kinds of languages, for example with free phrase order.

We think **Tily & Jaeger** are completely right to urge the use of ancillary evidence. Consider for example the role of village sign languages (Meir et al. 2010): unlike any spoken language, these are languages known to be of independent origin; spoken languages always have one or more ancestral donors, but the languages of the isolated deaf have to be constructed *de novo*. They provide crucial evidence for any supposed universal biases. If typology would clearly state its objectives in terms of testable hypotheses – not simply in terms of discovering world distributions of variants – it would come to have more fruitful interactions with the rest of the language sciences.

4. Other issues

4.1. *Invariance: The case of Bantu*

Our method for testing causal linkage between variables relies on change: when one value changes, the other should also. When faced with a very conservative language family like Bantu, whose time depth is relatively shallow, the method simply yields few results, because the method is sampling events – changes in word order – not languages with specific word orders. A number of commentators (e.g., **Baker**) point out that it would be very useful to extend the study to the encompassing Niger-Congo clade, and we completely agree. Some commentators (e.g., **Croft et al., Baker, Donohue**) wonder whether contact may actually cause such stasis, by providing neighbouring languages with the same

word orders. This is certainly possible, especially in multilingual settings (see, e.g., Gumperz & Wilson 1971), but it has no bearing on our method, which is a study based on change. We agree that the method at present allows little to be concluded in such cases of long-term stasis, although it is a useful reminder that syntactic patterns are capable of phylogenetic transmission over many thousands of years, whether or not this is buttressed by neighbouring languages. However, we do not agree that in general our approach lacks statistical power. As the Uto-Aztecan data in Figure 3 shows, even in relatively small families, we can detect functional dependencies (in this case a dependency between subject-verb and object-verb not predicted by Dryer's data), provided there is sufficient character state change.

4.2. *Variable rates of change*

It is a well-known feature of language change that rates are variable (e.g., Hoenigswald 1960, Nettle 1999, Blust 2000, Atkinson et al. 2008, Greenhill et al. 2010a). The apparent intractability of this problem is what led to the abandonment of phylogenetic inference by distance methods (lexicostatistics) by mainstream linguistics. Modern phylogenetics has made huge advances in this respect, and variation in rates of change is routinely included in phylogenetic models (Gray et al. 2009, Greenhill et al. 2010a). One method described by **Croft et al.** is the covarion. The covarion model (Penny et al. 2001) allows cognates (or other features) to change between faster and slower rates of change at different points on the tree. Our tree samples that underlie our phylogenetic hypothesis testing were all made using the covarion model as our previous work had shown that this model performed best in standard model comparison tests (Gray et al. 2009). **Croft et al.** suggest that we should also have used the covarion model when fitting the word order features onto the trees, hence enabling the features to change at faster and slower rates across the tree. We did not use this model in our original analyses as preliminary work showed that it made little improvement at the cost of far greater analysis run time and increase model parameterisation.

However, perhaps **Croft et al.** are right in suggesting that the covarion might be a better model of the evolution of word order features. We have run a single full cycle of pairwise tests in all four families using the covarion model (in the target article we repeated this ten times to ensure stability of the result). The preliminary results suggest that the covarion might improve the fit of the model to the data in some cases, but in more than 95 % of cases it does not. The inferred model parameters are similar to those reported in the target article. These preliminary results indicate that the covarion is, in fact, not a better model for the evolution of word order features. Finishing this analysis properly will, however, take several months. There is a final irony that we would like

to note: the covarion is a model parameter which allows one to better account for autocorrelated variation in rates of change. A typical example of autocorrelated variation in rates would be a feature that was highly variable on one branch of a tree, and invariant on another. This would in fact be an example of a lineage-specific process.

4.3. *Contact*

Many commentators feel that our study is flawed by ignoring contact (**Donohue, Bickel, Croft et al.**). It is true that contact could have effects on our basic vocabulary trees through lexical borrowing from other branches of the same family, and it could be the source of innovations in word order. But these processes are unlikely to vitiate our results for the following reason. First, with regards to the background tree construction, lexical borrowings from unrelated languages have no more impact than internal production of new lexical forms. But borrowings from related languages could, in principle, have an impact on the accuracy of the trees. Suppose our descendant languages B and C come into contact with a distant cousin, D: then D could in principle donate enough vocabulary to B and C to make them look like sisters. The central question is, therefore, what level of borrowing is required to substantially bias the tree estimates. Simulation analyses have shown that in most cases greater than 20% undetected borrowing is required to bias estimates of tree topology (Greenhill et al. 2009). In the construction of our trees known borrowings were removed from the analyses. Basic vocabulary generally has low rates of borrowing (Embleton 1986, Haspelmath & Tadmor 2009, Bowerman et al. 2011). Thus, while a small number of branches might be misplaced due to undetected borrowing, it is highly unlikely that our tree estimates are substantially biased by greater than 15% rates of undetected borrowing. In fact, as **Croft et al.** note, our tree estimates are quite congruent with the results of the comparative method.

Second, what about problems caused by the borrowing of the word order features between languages? Evolutionary processes propagate innovations across generations regardless of where they came from, whether generated by internal change (mutation) or external change (hybridisation). The transmission process is indifferent to the source of the innovation. What we were interested in was simply whether, IF there was a change in one variable, then did this induce a predicted change in another variable? It matters little whether the change in the original variable could be caused by internal or external processes. Nevertheless, further simulation studies we conducted showed that the comparative methods we used here are also robust to the effects of borrowing at this stage (Currie et al. 2010).

Under what conditions would contact seriously affect the findings? Suppose an ancestral language A has two descendants, B and C, that come into contact

with an unrelated language X. Suppose X donates a feature in the antecedent of an implicational universal (e.g., replaces VO with OV) – then to regain “harmony” one might expect the variant in the consequent to be induced (e.g., the change from Preposition to Postposition). If this occurs, our method will count the change as positive evidence in favour of the universal. So, clearly, “harmony” might be induced regardless of whether the change was internal or external. Now suppose X donates BOTH OV and Postposition in one blow. Now our method will mistakenly count this also as evidence for the universal. In other words the method may generate false positives for Greenbergian universals, because features might be borrowed together by chance, but it won’t deliver false negatives.

Readers can apply these remarks to the example raised by **Donohue**, namely the likely influence of SOV structures on Oceanic languages in the so-called Papuan Tip languages. These languages might have borrowed OV, then innovated Postpositions by harmony – which would be a strike for Greenberg, or they could have borrowed both together – which again we would have to count as one strike in favour of Greenberg. Despite this generosity to the rival hypothesis, we don’t find it strongly supported, as we reported.

In sum, contact effects exist, and are reflected in the geographical distribution of variants as pointed out by **Donohue, Bickel, Dryer** and other commentators, but they actually have little direct bearing on the problem at hand, namely, if innovations appear, due to either internal or external causes, do principles of Greenbergian harmony operate or not?

4.4. *Greenbergian principles*

Our methods tested for linkage between variables, capturing at least partially the Greenbergian notion of the tendency towards harmony. **Dryer** complains that we ignored the other Greenbergian principle of DOMINANCE – a simple preference for one value of one variable over another, for example a tendency for languages to prefer to order relative clause and noun as N-Rel rather than Rel-N, regardless of other word orders. It is true, that exploring dominance was not the focus of our endeavour. However, the methods employed can reveal such effects: transition probabilities to N-Rel regardless of other states can be measured. Future work should explore the possibility that a combined model of simple preferences plus implicational tendencies would fare better than the Greenbergian implications tested alone.

4.5. *Synchronic correlation vs. diachronic causal linkage*

Which of the following kinds of evidence provides better evidence for a causal linkage? (i) In a good sample of rooms, where a switch by the door is down, the light is on; (ii) In a good sample of cases, when the switch is moved down,

the light turns on. The latter evidence provides a temporal connection between the states, which is an additional precondition, above and beyond correlation, for a causal connection. Our study uses this additional criterion, and we should not be surprised that not all correlations turn out to be causally connected in this way.

Dryer's careful and useful commentary is nevertheless fixated on simple correlation, and he thinks we "consider a family to be one in which there is a correlation only if it is a correlating family in the sense just described". In fact, we are not in this same game at all. We are not classifying families, nor are we claiming, as he thinks, that Greenbergian word order correlations predict that the majority of families will conform. These are claims about the synchronic distribution of types across the languages of the world. We are asking a different sort of question: Can we find diachronic evidence for word orders in one part of the grammar to be linked to word orders in another part? That is, are the Greenbergian generalisations just statements about global distribution of types, or are they statements about causal linkages?

If Greenbergian generalisations act as actual causal biases, then they should be detectable again and again within a single lineage as it diversifies and evolves. Thus we are quantifying over instances of change, while **Dryer** quantifies over languages, genera, and continents. In his count, a whole genus can contribute just one data point on a claimed linkage between word orders; on our count, a single language can produce many data points as it evolves through time. His job is to detect the faint signal of correlations unexplained in the way most of them are by phylogeny and geography; our job is to detect biases in state transitions. **Bickel's** version is just another attempt to filter areal effects out and distill a cognitive residue. These more or less traditional typological studies are simply different enterprises from ours. Nevertheless, if we find no biases in state transitions, then that would remove one explanation for any synchronic correlation that **Dryer** (or **Bickel**) can find that is not already explained by phylogeny and geography.

Dryer's remarks should therefore be read as an interesting and important clarification of his own enterprise. Typologists will do well to study them carefully. His implicit message to us is: What is our rival explanation for these synchronic worldwide distributions? Our response is that we are not convinced that there are strong biases that cannot be accounted for by overt and covert phylogeny, contact, and the vagaries of history and demography. What **Dryer** tries to do is to deduct cases explicable by geography (contact) and phylogeny (common descent), and examine the residual pattern, assuming that there are no other explanations except cognitive biases. But, as we have already pointed out, controlling for ancient contact and unknown phylogeny is fraught with difficulties, and it is by no means obvious that the residue cannot be accounted by yet further factors. This is the central dilemma of classical typology.

We think our method offers a way out of the key dilemmas faced by traditional typology, and we urge readers to think the issues through.

5. The importance of phylogeny

Many of the commentators seem to imply that phylogeny (familial relatedness) is just not that important in typology. Hence the emphasis on areal or contact influences in the commentaries of **Dryer**, **Bickel**, **Donohue**, among others. We think the implicit downgrading of inheritance is misplaced for the following reasons. First, there is plenty of evidence that typological features are passed unchanged from generation to generation (see Dunn et al. 2005, Reesink et al. 2010, Greenhill et al. 2010a, Dediu 2011a, and the case of Bantu, already discussed) – the average rate of change for typological features is in the order of one change per n thousand years (where n is for many typological features in two or three figures). Linguists tend to slip from the demonstration that just about anything CAN be borrowed to the quite improbable assumption that a large proportion of structure actually is borrowed, forgetting that many areal effects are likely due to covert phylogenetic relatedness too deep to be detected by traditional methods. Second, even if structure is borrowed, it can be detected mostly because it is thereafter phylogenetically transferred and preserved. Third, treating phylogeny as the clear structural scaffold against which horizontal transmission can be detected, a strategy criticised in a number of the commentaries, is justified by the state of the art; we have a science of phylogenetic inheritance, but only the glimmerings of a science of lateral transmission (e.g. Gray et al. 2011, Nelson-Sathi et al. 2011).

Bickel, coming from a position close to the traditional typology concern with worldwide distributions, nevertheless (and to his credit) sees the importance of phylogeny. Like us, he wants to look inside language families for trends as a way – arguably the only systematic way – of controlling for shared inheritance. While **Dryer** pays attention to lineage only in order to partially neutralise it (by selecting only one language from each genus or high-level taxon), **Bickel** is much more interested in what happens within taxa. He takes the highest taxa (or maximal clades) within language families that have uniform VO ordering, and then examines preferences for GenN vs. NGen orders within each of his four major areas. From this he derives intrinsic preferences (Family Biases), and extracts those that hold up across areas. He uses 45 such taxa with 5+ languages and extrapolates to the rest of his sample, chiding us for our feeble use of only four taxa.

What **Bickel** doesn't seem to grasp is that our method quantifies over every reconstructed change within a whole family with hundreds of languages. Altogether we examined some 600 languages with their many ancestors, and generalised over hundreds of estimated changes of parameter values. Our method

doesn't work by simple inspection of synchronic bias – it works by reconstructing the most likely genealogy behind each family, so back-tracking along an extensive evolutionary trail. In tracing through every branch of each of these language families, we have traversed all the changes in our selected major phrases in about half a million years of language evolution – for each of these branches is an independent diachronic experiment. We have not examined a mere four taxa: we have examined hundreds of changes, and systematically tested for biases caused by implicational word order universals.

Bickel's Family Bias Method is, he freely admits, rather crude. His approach is actually remarkably similar to a much earlier approach in biological phylogenetics, the method of Independent Contrasts (Felsenstein 1985), now largely superseded by the approach we have used. In contrast, the methods we employed have been in extensive use in biological samples for years, where they work better than cross-species sampling for discovering linked genes (Pagel 1994, 1999), just as they offer better ways of testing dependencies between typological variables in language materials.

In short, **Bickel** adopts our tactic of controlling for relatedness by looking within language families. But he has not grasped that typological generalisations can be tested not just by counting languages and taxa, but by statistically analysing hundreds of diachronic events within a single family (here again, see **Cysouw** for the logic).

6. Conclusions

These discussions throw into stark relief the question: What is typology for? Does it have theoretical ambitions, and if so, what are they? Possible answers include:

- (i) Typology is meant to reveal underlying universal principles of cognitive processing of language, or universal functional pressures. Our comment: For this goal, our method for finding correlated changes is much more revealing, and less subject to Galton's problem.
- (ii) Typology reveals the limitations on the design space for possible languages. Our comment: Given the slow time scale of language change, and the relatively short period (about 70,000 years) since the diaspora of modern humans, it is unlikely that the current variation reflects broad sampling of the possibility space (see Evans & Levinson 2009). The methods required to explore these issues (e.g., massive simulations of language diversification) also lie in the bioinformatics domain.
- (iii) Typology is meant to reveal historical processes of language change and language contact. Our comment: Here, phylogenetic processes, and computational methods for investigating hybridisation, are likely to be much more revealing than a bunch of data points with geographic coordinates (even though those are required data).

- (iv) Typology has a basically descriptive function: its job is to outline the scope and types of linguistic diversity, without theoretical commitment to any particular enterprise. Our comment: This is a hugely valuable enterprise that is maid servant to many different lines of enquiry. Our own enquiries would obviously have been impossible without the industry of a relatively small number of dedicated typologists (many represented by the commentaries here) who have painstakingly compiled data from many different sources.

We believe that if typology is to regain the theoretical high ground, it must embrace the new methods that are coming from bioinformatics and evolutionary biology. Only then will it be in a position to answer the leading questions about the origin and diversification of languages, the effects of underlying biases whether cognitive or systemic, and provide evidence for constraints on change. A typology so revitalised would be a natural partner for the other sciences interested in human evolution, cognition, and diversity.

Received: 11 October 2011

Max Planck Institute for Psycholinguistics
University of Auckland

Correspondence addresses: (Levinson, corresponding author) Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, The Netherlands; Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands; e-mail: stephen.levinson@mpi.nl. (Greenhill) Department of Psychology, University of Auckland, Auckland 1142, New Zealand; Computational Evolution Group, University of Auckland, Auckland 1142, New Zealand; e-mail: s.greenhill@auckland.ac.nz. (Gray) Department of Psychology, University of Auckland, Auckland 1142, New Zealand; e-mail: rd.gray@auckland.ac.nz. (Dunn) Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, The Netherlands; e-mail: michael.dunn@mpi.nl.

Acknowledgements: We would like to thank Frans Plank for giving us this opportunity to engage in this public discussion of issues that we believe are important to the future of linguistic typology. Special thanks to Fiona Jordan for detailed suggestions on an earlier draft.

References

- Amos, William & Joe I. Hoffman. 2009. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society B: Biological Sciences* 1678. 131–137.
- Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346–349.
- Atkinson, Quentin D., Andrew Meade, Chris Venditti, Simon J. Greenhill & Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science* 319. 588.
- Blust, Robert. 2000. Why lexicostatistics doesn't work: The "universal constant" hypothesis and the Austronesian languages. In Colin Renfrew, April M. S. McMahon & Robert L. Trask (eds.) *Time depth in historical linguistics*, Vol. 2, 311–331. Cambridge: McDonald Institute for Archaeological Research.

- Bouckaert, Remco. 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26. 1372–1373.
- Bowern, Claire, Patience Epps, Russell Gray, Jane Hill, Keith Hunley, Patrick McConvell & Jason Zentz. 2011. Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS ONE* 6(9). e25195. <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0025195>
- Currie, Thomas E., Simon J. Greenhill & Ruth Mace. 2010. Is horizontal transmission really a problem for phylogenetic comparative methods? A simulation study using continuous cultural traits. *Proceedings of the Royal Society B: Biological Sciences* 365. 3903–3912.
- Dediu, Daniel. 2011a. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B: Biological Sciences* 278. 474–479.
- Dediu, Daniel. 2011b. Are languages really independent from genes? If not, what would a genetic bias affecting language diversity look like? *Human Biology* 83. 279–296.
- Drummond, Alexei J. & Andrew Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7. 214. <http://www.biomedcentral.com/1471-2148/7/214>
- Dryer, Matthew S. 2008a. Order of subject and verb. In Haspelmath et al. (eds.) 2008, Chapter 82. <http://wals.info/feature/82>
- Dryer, Matthew S. 2008b. Order of object and verb. In Haspelmath et al. (eds.) 2008, Chapter 83. <http://wals.info/feature/83>.
- Dryer, Matthew S. 2008c. Order of adposition and noun phrase. In Haspelmath et al. (eds.) 2008, Chapter 85. <http://wals.info/feature/85>.
- Dryer, Matthew S. 2008d. Order of genitive and noun. In Haspelmath et al. (eds.) 2008, Chapter 86. <http://wals.info/feature/86>.
- Dryer, Matthew S. 2008e. Order of adjective and noun. In Haspelmath et al. (eds.) 2008, Chapter 87. <http://wals.info/feature/87>.
- Dryer, Matthew S. 2008f. Order of demonstrative and noun. In Haspelmath et al. (eds.) 2008, Chapter 88. <http://wals.info/feature/88>.
- Dryer, Matthew S. (2008g) Order of numeral and noun. In Haspelmath et al. (eds.) 2008, Chapter 89. <http://wals.info/feature/89>.
- Dryer, Matthew S. 2008h. Order of relative clause and noun. In Haspelmath et al. (eds.) 2008, Chapter 90. <http://wals.info/feature/90>.
- Dunn, Michael. 2009. Contact and phylogeny in Island Melanesia. *Lingua* 119. 1664–1678.
- Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson & Neele Becker. 2011b. Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronica* 28. 291–323.
- Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011a. 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 & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84. 710–759.
- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309. 2072–2075.
- Embleton, Sheila M. 1986. *Statistics in historical linguistics*. Bochum: Brockmeyer.
- Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32. 429–448.
- Felsenstein, Joseph. 1978. The number of evolutionary trees. *Systematic Zoology* 27. 27–33.
- Felsenstein, Joseph. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17. 368–376.

- Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *American Naturalist* 125. 1–15.
- Felsenstein, Joseph. 1988. Phylogenies and quantitative characters. *Annual Review Ecology and Systematics* 19. 445–471.
- Felsenstein, Joseph. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer.
- Gray, Russell D., Quentin D. Atkinson & Simon J. Greenhill. 2011. Language evolution and human history: What a difference a date makes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366. 1090–1100.
- Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323. 479–483.
- Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405. 1052–1055.
- Greenhill, Simon J., Quentin D. Atkinson, Andrew Meade & Russell D. Gray. 2010a. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences* 277. 2443–2450.
- Greenhill, Simon J., Thomas E. Currie & Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B: Biological Sciences* 276. 2299–2306.
- Greenhill, Simon J., Alexei J. Drummond & Russell D. Gray. 2010b. How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PLoS ONE* 5(3). e9573. <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0009573>
- Greenhill, Simon J. & Russell D. Gray. 2009. Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. In Karl A. Adelaar & Andrew Pawley (eds.), *Austronesian historical linguistics and culture history: A festschrift for Robert Blust* (Pacific Linguistics), 375–397. Canberra: Australian National University.
- Gumperz, John J. & Robert Wilson. 1971. Convergence and creolization: A Case from the Indo-Aryan/Dravidian border in India. In Dell H. Hymes (ed.), *Pidginization and creolization of languages*, 151–167. Cambridge: Cambridge University Press.
- Harvey, Paul H. & Mark D. Pagel. 1991. *The comparative method in evolutionary biology*. Oxford: Oxford University Press.
- Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86. 663–687.
- Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2008. *The world atlas of language structures online*. München: Max Planck Digital Library. <http://2008.wals.info>
- Haspelmath, Martin & Uri Tadmor (eds.). 2009. *Loanwords in the world's languages: A comparative handbook*. Berlin: De Gruyter Mouton.
- Hennig, Willi. 1966. *Phylogenetic systematics*. Urbana, IL: University of Illinois Press.
- Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33. 61–83.
- Hoeningswald, Henry M. 1960. *Language change and linguistic reconstruction*. Chicago: University of Chicago Press.
- Huelsenbeck, John P., Bruce Rannala & John P. Masly. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288. 2349–2350.
- Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen & Jonathan Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294. 2310–2314.
- Kari, James & Ben A. Potter (eds). 2010. *The Dene-Yeniseian connection* (Anthropological Papers of the University of Alaska, New Series 5.) Fairbanks, AL: Department of Anthropology, University of Alaska, Fairbanks.
- Kass, Robert E. & Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430). 773–795.
- Kirby, Simon, Mike Dowman & Thomas Griffiths. 2007. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences* 104. 5241–5245.

- McMahon, April M. S. & Robert McMahon. 2005. *Language classification by numbers*. Oxford: Oxford University Press.
- Mace, Ruth & Mark Pagel. 1994. The comparative method in anthropology. *Current Anthropology* 35. 549–564.
- Meir, Irit, Wendy Sandler, Carol Padden & Mark Aronoff. 2010. Emerging sign languages. In Marc Marschark & Patricia Spencer (eds.), *Oxford handbook of deaf studies, language, and education*, Vol. 2, 267–280. Oxford: Oxford University Press.
- Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin & Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences* 278. 1794–1803.
- Nettle, Daniel. 1999. Is the rate of linguistic change constant? *Lingua* 108. 119–136.
- Pagel, Mark. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences* 255. 37–45.
- Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401. 877–884.
- Penny, David, Bennet J. McComish, Michael A. Charleston & Michael D. Hendy. 2001. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *Journal of Molecular Evolution* 53. 711–723.
- Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7(11). e1000241. <http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1000241>
- Steel, Mike & David Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* 17. 839–850.