1. Introduction

In their paper “Evolved structure of language shows lineage-specific trends in word order universals”, Dunn, Greenhill, Levinson and Gray present evidence purporting to demonstrate that both Chomskyan and Greenbergian language universals are invalid. In particular, and of most interest to readers of this journal, they state “contrary to the Greenbergian generalizations, we show that most observed functional dependencies between traits are lineage-specific rather than universal tendencies” (Dunn et al. 2011:79). If this conclusion were correct, the field of typology would have to change profoundly: Greenbergian universals would no longer exist, and the correlations that typologists have attempted to explain in terms of semantics, discourse, processing and other general cognitive or interactional terms would have to be explained in “culture-specific” terms. This conclusion was taken up in the general media as well as in a number of linguistics electronic discussion lists.
Dunn et al.’s analysis merits close attention, for several reasons. Although the method they apply is quite different from the method used by typologists to derive the Greenbergian universals in the first place, Dunn et al.’s method is one that many typologists from Greenberg onward have aimed for. Also, although Dunn et al. used statistical modeling methods that are unfamiliar to typologists and difficult to interpret for someone lacking a statistical background, these methods hold the promise of allowing for significant progress in typology. We hope that our comment will suggest ways for a typologist to evaluate statistical analyses such as Dunn et al.’s.

We argue in this comment that certain assumptions made by Dunn and colleagues in the application of the model pose serious issues in accepting the conclusions, notably the absence of any Type II error analysis to assess the rate of false negatives, the absence of contact effects and the nature of the phylogenies used. Although our examination of Dunn et al.’s analysis is critical, we nevertheless believe that typologists should welcome the model and encourage the development of a revised model with more linguistically plausible assumptions.

2. Synchronic and diachronic approaches to word order universals

Greenberg (Greenberg 1966a) and his successors derived the Greenbergian word order universals from a synchronic sample of languages. The data from the synchronic sample justified the positing of implicational and (rarely) biconditional universals of word order such as “If a language has object-verb order, then it also has subject-verb order”, or “If a language has prepositions, then it has verb-object order, and if it has postpositions, then it
has object-verb order” (for a statistical technique to determine whether implicational universals with exceptions are justifiable, see Maslova 2003).

It is well known among typologists that the synchronic method has methodological problems. The first problem is that the cases in the linguistic sample may not be independent, or more precisely, the sample of languages whose word order distribution motivates the Greenbergian universals may not be typical of the set of languages it is intended to represent, that is the set of human languages. Much has been written about this problem in typology (inter alia, Bell 1978, Perkins 1989, Dryer 1989a, Rijkhoff and Bakker 1993). The usual approach to address this problem in typology has been to construct samples stratified by genetic family and geographical area, since it is known that these two factors strongly influence the typological traits of languages. Another approach to this problem is to explicitly include these factors in a statistical analysis. Doing so explicitly identifies the contribution of these factors to the current distribution of traits, and allows one to determine if any of the observed distribution can be attributed to a correlation of traits independent of those factors. Atkinson (2011) takes this approach in controlling for genetic family in order to examine postulated correlations between a measure of phonological complexity, speech community population size, and distance from a possible origin point in Africa (see Jaeger et al., this issue; Maddieson et al., this issue).

Nevertheless, the problem of typicality of a sample is difficult if not impossible to resolve. Every sample will always include deviations from the typical behavior; the important question is whether these deviations can be misinterpreted in analysis to create systematic errors. There is no obvious way to establish that one has a typical sample, and
in finite systems, this question cannot even be made precise. We thus acknowledge that every statistical analysis carries an implicit uniformitarian precondition supposing that the sample is typical of the population of interest (i.e. whether the cases in the sample were independently drawn from the population or at least drawn under conditional independence, cf. Jaeger, this issue). If this condition is not met the conclusions of the statistical analysis are not valid. Therefore, beyond statistical tests within a model, full model validation continues to test the representativeness of the samples.

There is a second problem with synchronic language samples, namely the causal model that is taken to underlie the synchronic distribution. The causal model assumed by most typologists is an ultimately diachronic one (see, inter alia, Greenberg 1966b, 1969; Bybee 1988; Maslova 2000; Croft 2003, chapter 8). For example, a correlation of word orders observed in a synchronic language sample is presumed to be the result of diachronic processes so that a change in one word order will eventually cause a change in the correlated word order. One way to use the synchronic data to derive properties of the diachronic process is to assume as a precondition that the current synchronic distribution of typological traits is a stationary distribution. That is, the current distribution does not reflect any effect from the initial state of the system—the typological traits possessed by the original human protolanguage that is the ancestor of all contemporary human languages (or protolanguages, if one believes in linguistic polygenesis). There are occasional suggestions that not all typological traits exhibit a stationary distribution (e.g. Maslova and Nikitina 2010).

An approach to causal modeling that does not require this stationarity assumption is to directly fit a state-process model of the change of typological traits to data in their
historical context. By modeling the relevant aspects of the process correctly, the redundancy in synchronic samples due to relatedness is correctly weighted, and the processes that are likely to lead to the observed, possibly non-stationary, distribution are identified. This approach has been advocated by typologists from Greenberg (1978) to Maslova (2000); see also Croft (2003, ch. 8; 2007): This approach is what Greenberg calls a dynamicization of a synchronic typology (Greenberg 1969:75): use synchronic variation in distribution to hypothesize processes of language change leading to the synchronic distribution. This is the approach taken by Dunn et al. In other words, the method that Dunn et al. borrow from Pagel and Meade is a realization of one of the goals advocated by many typologists for over four decades.

3. A state-process model for diachronic typology

3.1. The Pagel and Meade trait evolution model

Dunn et al. adapted a state-process model of trait evolution originally developed to investigate whether pairs of phenotypic traits evolved in an independent way, or in a dependent or contingent way (Pagel, 1994; Pagel & Meade, 2006). This model, named Discrete, is freely available in the software package BayesTraits, and describes changes in traits over branches of a phylogenetic tree via a continuous-time Markov process. The independence of two traits can be assessed in such a framework in two ways. First, it is possible to directly compare how well models of independent evolution versus dependent evolution describe the available data (e.g. by means of the Bayes factor, Jeffreys, 1935;
for an introduction see Kass and Raftery, 1995). This is the approach taken by Dunn and colleagues. An alternative method of comparison is to sample from the distribution of models implied by the data—some of which will describe dependent and some independent trait evolution—using a so-called reversible-jump method (see below).

Consider a trait A, which can take one of two state, denoted 0 and 1. For example, trait A might describe the ordering of adjectives and nouns with value 0 for adjective-noun order and value 1 for noun-adjective order. Then, let $q_{01}$ be the instantaneous rate of change from state 0 to state 1, such that the probability that a species which starts with $A=0$ will change to $A=1$ after infinitesimal time $dt$ is $dt \times q_{01}$ (Figure 1). The probability of no change is thus $1 - dt \times q_{01}$. Based on these instantaneous rates, it is possible to derive the corresponding transition probabilities after a finite length of time $t$, taking into account the fact that the state may flip-flop many times in that interval.¹

**FIGURE 1 AROUND HERE**

By using such a continuous-time model, and representing state changes using instantaneous rates of change, it is possible to describe state changes over a phylogenetic tree, where contemporary language traits are observed (or prehistorical traits are inferred) only at particular points, separated by variable lengths of time. Furthermore, because transitions are directly modeled via probabilities, it is reasonably straightforward to use standard statistical methods to derive estimates of the rates which take into account uncertainty about the actual phylogenetic tree (which is not known for certain but has

¹ Technically, this is done by taking the matrix exponential of the instantaneous rate matrix scaled by $t$ (Pagel, 1994).
been reconstructed on the basis of genetic material or cognate lists) and the ancestral states which must be inferred on the basis of known states and the rates of change.

3.2. Assessing dependence between two traits

Now consider a pair of traits, A and B, each of which can be 0 or 1. There are four possible state pairs \{A,B\}: state 1={0,0}, 2={0,1}, 3={1,0}, and 4={1,1}. State transitions are modeled in exactly the same way as before when there was only a single trait, with one exception: it is commonly assumed that instantaneous rates of changes that correspond to a simultaneous change in both traits, such as \(q_{14}\), are defined to be zero (Figure 2). This is done to assign a vanishing probability to truly simultaneous change in two traits, and inclusion of such additional terms are usually not useful. In the two-trait model, then, there are eight rates which are used to describe the data, corresponding to all 12 possible state changes minus the four where both traits change simultaneously (\(q_{14}\), \(q_{41}\), \(q_{23}\), and \(q_{32}\)).

FIGURE 2 AROUND HERE

Independent and dependent evolution of traits A and B imply different structures on these rates. If traits A and B evolve independently of each other, then the rate at which A changes from 0 to 1 should not depend on the value of B, and thus \(q_{13} = q_{24}\) (Figure 2, red arrow). The same holds for the other three symmetric pairs: \(q_{12} = q_{34}\), \(q_{31} = q_{42}\), and \(q_{21} = q_{43}\). Thus, a model of independent evolution has at most four unique rate parameters (Figure
3). On the contrary, if any of the above equalities are broken, then the rate of change of one trait depends on the current state of the other in at least one case, with the most extreme possibility being that every one of the eight rates is allowed to vary freely.

FIGURE 3 AROUND HERE

Given these two types of models, there are a variety of ways to assess whether the additional complexity of the dependent evolution model is warranted given the data. The two methods discussed here rely on drawing a large number of samples from the posterior distribution of rates in each model given the data, using Markov Chain Monte Carlo (MCMC) methods. MCMC sampling is standardly used in modern statistics to test hypotheses where parametric distributions cannot reasonably assumed.

The first, and the method used by Dunn et al., is to calculate the Bayes Factor of the dependent versus the independent model, which can be estimated based on the MCMC sample of each model. The Bayes Factor is similar to a likelihood ratio test. However, instead of using the maximum likelihood of the data under the best-fitting parameters for each model, the marginal likelihood of the data under each model is used. The marginal likelihood of the data under a given model is computed by averaging the likelihood of the data over all possible parameter values, weighted by their prior probability. This quantity is also called the evidence of a model, since it corresponds to the degree to which the model's entire range of predictions match the observed data. The evidence has a geometrical interpretation, since it is roughly equivalent to the maximum likelihood of the data, multiplied by the proportion of the volume of the model's parameter space.
where the model's predictions match the data reasonably well (MacKay, 2003:343). This penalizes unnecessarily complex models, whose predictions match the data very well for some parameter values but are qualitatively different for most others. Thus, when comparing a complex model to a simpler model, a high Bayes Factor (the log-ratio of the evidence of each model) indicates that the additional complexity of the complex model is warranted by the data.

In the case of models of trait evolution, models of dependent evolution are, in general, more complex, since each of the eight rate parameters may vary freely while in an independent model there are only four freely varying rate parameters. However, there are many specific models of dependent evolution that have fewer than eight rates, and some with fewer than four. The only qualification a model must meet in order to be dependent is for one of the four equations above to not hold; for an extreme example, consider the model where $q_{13} = r_1$, while the other seven rates are $r_2$. This model is dependent, since the rate for A=0 becoming A=1 when B=0 ($q_{13}=r_1$) is different from the rate for the same change in A when B=1 ($q_{24} = r_2$). In fact, most implicational, and many biconditional, correlations could arise from such simpler models that do not need all eight parameters. This fact is highly relevant to evaluating Dunn et al.'s (2011) results because a failure to accept the fully dependent model over the simpler independent model indicates only that the full complexity of the eight-parameter dependent model is not justified by the data, and does not rule out more restricted but dependent model.

A second approach to comparing the dependent and independent models which potentially avoids some of these issues is to actually sample models in addition to the rates of trait evolution, using a reversible-jump MCMC scheme (Pagel & Meade, 2006).
Here a model is considered more generally than simply dependent versus independent, and is any grouping of rates. The simplest possible model is one where all eight rates are exactly equal, and there is only one free parameters; the most complex is, again, the model where all eight rates are allow to vary freely. While this scheme is more flexible in that it allows to go beyond the simple hypothesis, it is not the method that Dunn et al. used.

4. Assessing Dunn et al.’s results in light of how they analyzed the data

Having described what Dunn et al. actually did to obtain their results, we may now turn to the question of assessing their statistical procedure. How can a typologist (or a historical linguist) interpret the results based on the model Dunn et al. applied, and indirectly assess the utility of the model that Dunn et al. used? A typologist or historical linguist looking at a phylogenetic model applied to typological data would have three questions they would want answered:

1. What are the assumptions behind the model that gave these results? How empirically problematic are the assumptions? Any model must make simplifying assumptions in order to make the model tractable, and in order to provide a meaningful analysis constrained by the available empirical data. Some of these assumptions may not be widely accepted among linguists, and some assumptions may even be contrary to what is widely accepted. The first question is to determine exactly what the assumptions are
behind the model, and compare them to what is empirically supported, or at least what is generally believed, among linguists.

2. How sensitive are the results to the assumptions? That is, how likely is it that the results would be different if different assumptions were made (in particular, different assumptions that are closer to what is considered plausible)? Some assumptions may be major factors in bringing about the results obtained: changing those assumptions is likely to significantly change the results. Other assumptions, even if they are empirically implausible, may not actually play a major role in bringing about the results obtained: changing those assumptions may not significantly change the results. In other words, not every implausible assumption of the model may be grounds for rejecting the results. Of course, some assumptions are easier to change in the model than others. This brings us to the last question.

3. What would it take to run and test a model with more plausible assumptions (and thus, to be able to answer 2 with greater precision)? Would it take revising the current model, or developing a new model? Would it take a larger set of empirical data? The third question is relevant, both for answering question 2, and for judging what is reasonable or unreasonable to expect someone like Dunn et al. to take into consideration. Changing some assumptions, important as they may be, may require a large amount of computing time or a major investment in redesigning the software, or a major investment in collecting additional data. For example, Dunn et al. used cognate data for the phylogeny construction that was obtained in collaboration with language family experts over a long
period of time, and word order data from WALS that was collected by Matthew Dryer over a long period of time, and then ran them through off-the-shelf software programs—BayesPhylogenies and BayesTraits; they did not develop new mathematical models/software. If only minor changes are require to accommodate more plausible assumptions, then the model should be useful to typologists even if the results reported in Dunn et al. (2011) are suspect due to the assumptions they made. If, on the other hand, major revisions to the model are required to change assumptions that would have a major effect on the results, then typologists are more likely to wait and see until a better model is forthcoming.

Before examining the assumptions in Dunn et al. and attempting to answers questions 1-3 for each assumption, we must clarify exactly what are the results of Dunn et al. that are of interest to typologists. There are actually two results to consider:

(i) The presence/absence of the 28 pairwise correlations in one or more of the 4 phylogenies they tested;

(ii) The conclusion, based on the non-uniformity of (i), that word-order correlations are lineage-specific.

Even just one non-uniform result in (i) provides some support to (ii), at least with respect to the word order correlation with the non-uniform result. But if there is reason to question the validity of (i) in general, then (ii) is also questionable.
The assumptions of Dunn et al.’s model with respect to the phylogenetic and typological data they use are listed in (1):

(1) a. The synchronic distribution of word orders is the result of diachronic transmission processes.
b. Accurate phylogenies for individual families can be constructed using only presence/absence of genuine cognates in wordlists as short as 92 words (that being the length of the shortest wordlist used in Dunn et al.’s study).
c. The choice of a cut-off value for the Bayes Factor they propose (i.e., BF=5) is an arbitrary but conventional value attempting to balance Type I (false positive, i.e., mistaking random fluctuations as evidence of correlation) and Type II (false negative, i.e., not detecting correlations because they do not meet the threshold of the test) error rates.
d. Asymmetric and symmetric correlations, corresponding to one-way implicational and bi-directional universals, respectively, were lumped together (Dunn et al. tested only dependent vs. independent models).
e. Languages with “no dominant word order” were coded as polymorphic.
f. Word order properties were compared pairwise.
g. Typological traits are exclusively co-inherited with lexical traits, i.e. no "horizontal transmission" or contact.
h. Some kind of constancy in the trait evolution process is assumed.
j. Depending in part on assumptions in (h), Dunn et al. may be taken to assume polygenesis; see §4.4 for further discussion.
Next, we discuss these assumptions in turn.

4.1. Assumptions (1a) and (1b): Diachrony and phylogeny

The first assumption, that the synchronic distribution of word orders is the result of diachronic transmission processes, is the foundation of Dunn et al.’s model. As discussed in §2, this assumption of the model is widely shared among typologists. It is of course not the only logically possible hypothesis. For example, one could hypothesize that children construct the typological traits of their language de novo in each generation. Empirically, however, the evidence is overwhelming that children’s language, including its typological traits, are acquired on the basis of the traits of the language spoken around them. We make this point explicit as a reminder of why Dunn et al.’s model should be of interest to typologists; the foundational assumption behind both Dunn et al. (2011) and the theory of typological distributions held by many typologists is essentially the same. For this reason, we do not have to consider how shifting to a model lacking this assumption might affect the results. We note that Dunn et al. use a very schematic description of the language states—one order of a construction vs. its opposite—and so the state-process model developed by them does not take into consideration other grammatical details of the constructions which may be relevant to the mechanisms of the change process. Nevertheless, a schematic state-process model of word order change provides a useful starting point for exploring the specific mechanisms that may bring about constructional change.
The second assumption is that accurate phylogenies of the individual families can be reconstructed using only presence/absence of genuine cognates in wordlists, sometimes as short as 92 words. In fact, most historical linguists would consider this assumption to be problematic. In historical linguistics, phylogenies are constructed based not only on presence/absence of cognates, but also on sound correspondences and the regular sound changes those correspondences are evidence for, as well as morphological processes that can be reconstructed in the family.

How much of a difference does changing this assumption make? Atkinson et al. (2005) use phonological and morphological traits for Indo-European from Ringe et al. (2002) in a similar phylogeny reconstruction algorithm to that used in Dunn et al., and conclude that the resulting trees are very similar to those reconstructed using presence/absence of cognates. Also, Dunn et al. use 600 trees in the posterior sample of trees (4200 trees for Austronesian; Dunn et al. 2011S:2), so that their trait evolution analysis is not dependent on the peculiarities of any one tree, such as the trees used for illustration in the article and the supplementary materials. Based on these observations, we conclude that it is unlikely that adding phonological and morphological evidence to the phylogeny reconstruction is going to significantly change the results of the trait evolution model, though by reducing the uncertainty in trees, they may increase the power to uncover weaker correlations. Moreover, lacking a protolanguage reconstruction algorithm that provides a statistical measure of uncertainty on the inferred history, it is not presently possible to improve on this assumption without a major effort.

4.2. Assumptions (1c) to (1f): Power and related issues
The third assumption is that choice of a cut-off value for the Bayes Factor they propose is an adequate balance between Type I and Type II error rates if the assumptions of the model are met. But empirically, it is very obvious that there are Type II errors in the data. At the very least, for 18 of the 28 word order pairs examined, there are no word order changes at all in the Bantu tree. This fact is alluded to only in the caption of Figure 2 in the main article. Hence, there is clearly insufficient evidence to detect 18 of the 28 word order correlations in Bantu. Visual inspection of the trees presented in the paper and the supplementary materials suggest that in at least some cases where there are some word order changes in the tree, the changes occur sufficiently rarely that estimating the Type II error rate is certainly warranted.

These observations indicate that in fact, the Type II error rate may be sufficiently large to require a reinterpretation of the results. In particular, result (i) may change so that many of the non-correlations reported by Dunn et al. in result (i) are due simply to absence of sufficient evidence—enough that the generality of result (ii) is questioned.

What would it take to test for Type II errors? A full power study of the data would be quite difficult. However, since the basic result of the article crucially depends on this calculation, some sort of testing should have been done. There are alternative plausibility arguments Dunn et al. could have tried without doing a full power calculation. For example, one could check if a likelihood ratio test would have had enough power for some small sample of credible trees (not difficult with BayesTraits). One could also monitor the posterior distribution of the correlated change matrices they obtained to have an idea of the width of the distribution. They could even have seeded the Bantu run with
the same amount of correlation they saw in some deeper tree and monitored its trajectory to guess whether this was outside the valley of preferred matrices. Whatever method is chosen, some testing of whether there is enough power in the data to justify the results is required.

Given the sparseness of the word order variation in the data, implying few word order change events in the families examined by Dunn et al., we anticipate that power simulations would change the interpretation of the results from evidence of absence of a correlation—what is claimed by Dunn et al.—to absence of evidence for a correlation. This is, obviously, a rather crucial difference. The question then is, what would it take to obtain enough data to determine whether or not the correlation in fact held in a language family? This is in fact a difficult question to answer. The problem is that the language families that are widely accepted by linguists are quite shallow relative to the history of human language, and more pertinent to the question at hand, quite shallow relative to the rate of word order change. Word orders are relatively stable in most language families for which we have good phylogenies and large numbers of languages. Adding more Bantu or Austronesian languages to the datasets used by Dunn et al. is unlikely to change that.

Only discovering deeper phylogenies will bring in data that could answer this question. The problem is that historical linguists are pessimistic, or at least lack a consensus, that this is even possible for currently accepted language families. In fact, for Bantu one could look at Niger-Congo, of which Bantu is universally considered to be a subgroup. (Dunn et al. presumably did not do this because of the absence of verified cognate sets for Niger-Congo; but they do not mention this.) If we are restricted to the shallow phylogenies that are available at present, one may not be able to determine
whether or not a word order correlation holds in a language family. (For example, one
cannot do it for the many families in the WALS genetic classification that have only one
language; see also Tily and Jaeger (this issue) who take this issue to argue that one can
complement typological approach with emerging behavioral methods to test the validity
of linguistics universals).

Assumptions (d), (e) and (f) are all problematic from an empirical perspective, but all
are related to the power issue. Dunn et al. test only two alternative trait evolution models,
an INDEPENDENT one in which there is no correlation between the paired word orders, and
a DEPENDENT one in which any type of correlation between the paired word orders is
assumed. This choice means that asymmetric correlations (i.e. one-way implicational
universals) and symmetric correlations (i.e. biconditional universals) are lumped together.
Empirically, the vast majority of word order universals (and, in fact, of typological
universals in general), are asymmetric; there are hardly any symmetric universals.

Lumping asymmetric and symmetric universals will affect the results in at least some
cases, when there are a small number of change events in the sample. While the method
used in Dunn et al. should still return valid statistic in a sufficiently large sample of
events, when the sample size is limiting, a reversible-jump MCMC may reveal a
preference for a submodel even when the more inclusive (and complex) model is not
required by the data. Apart from this methodological issue, the power to detect an
asymmetric model over the null model of independence is often smaller: So, it may not
be rated highly enough in small data sets, and overall the null hypothesis (the
independent model) will be judged adequate. For example, in the subject-verb/object-
verb pair, subject-verb order is very heavily preferred over verb-subject order. Although
synchronic typological studies indicate an asymmetric dependency between subject-verb and object-verb order, it may not be preferred over the independent model unless the data set is large enough to see enough changes of the relevant kinds.

Although changing this assumption may not change the results in too many cases, it is in fact possible in BayesTraits to treat asymmetric and symmetric dependency models separately. Since the number of word order changes in the data is indeed small this modification seems worth making.

Assumption (e) pertains to the treatment of word orders that Dryer coded as “no dominant order”. Dunn et al. selectively quote Dryer as indicating that “no dominant order” in some languages indicates a genuinely polymorphic state: one order is dominant in a subset of forms or constructions, and the other order is dominant in another subset. In fact, the passage quoted also indicates that word order was coded as “no dominant order” when both orders occur variably, but one is not much more frequent than the other (see Dryer 1989b). In this case, there may be a minimal difference between coding a language as having no dominant order, and a language having a dominant order—only a difference of a few percentage points. But in fact the majority of cases of “no dominant order” are cases in which two orders are mentioned in the language documentation, but no further information is given that would allow Dryer to determine which order, if any, were dominant (see Dryer, this issue). Hence the assumption that “no dominant order” indicates a polymorphic order is an unwarranted assumption.

Dryer (this issue) suggests removing such languages from the analysis, hence effectively treating them as missing data. Recoding these data is of course straightforward to do. The number of examples coded as polymorphic is relatively small,
so it would not be expected that the results would change significantly. However, removing the polymorphic traits would also reduce the overall sample of changes, and therefore aggravate the power problem described above.

Assumption (f) was that Dunn et al. examine all possible pairwise correlations between the word orders in their data. This is also the practice of many typologists. However, it may give rise to spurious word order correlations (i.e. it inflates the Type I error): A may correlate with B, but this may be due to a correlation between A and C and between C and B. Or there may be a higher-order correlation, say a three-way correlation between A, B and C, that is missed by individual pairwise correlations. Justeson and Stephens (1990) use a log-linear analysis of synchronic data that simultaneously compares all word orders in order to tease out the correlations (pairwise or higher-order) that give the best fit to the data. They found that some but not all pairwise word order correlations previously proposed in the typological literature were supported. This issue can be addressed by more sophisticated analysis which acts on all the word orders simultaneously, such as forward selection and backward elimination of interaction terms to the null model without interactions (but see Harrell 2001:56ff. for the dangers of forward selection and backward elimination). It is, however, unlikely that this change in itself would significantly change the results of Dunn et al.’s analysis.

In sum, the power issue is a serious one in that it is clearly a problem with this dataset, and addressing it is likely to significantly change the results. The other assumptions mentioned here are either aggravated by the power issue (asymmetric vs. symmetric correlations; simultaneous comparing all word order changes) or would
further aggravate it (treating “no dominant order” languages as missing data rather than as polymorphic types).

Nevertheless, it appears that not every absence of a correlation is due to insufficient data. For example, the absence of a correlation between adposition-noun and verb-object order in the Uto-Aztecan phylogeny presented in the article appears likely to be significant, based on a visual inspection of the tree and the data presented in the article.

Evaluation by visual inspection can be done as follows:

1) Mentally reconstruct the ancestors weighting the descendants in inverse proportion of the branch lengths (i.e. closer descendants count more). Overall uncertainty in reconstruction depends on the rate of the trait and its directionality (faster traits are reconstructed badly, highly directional traits are reconstructed well). Rate is estimated from the tree overall: faster traits should be randomized more. Directionality, too, is estimated from the tree overall: highly directional traits will have fewer A within B than B within A clusters than what one would simply expect based on the overall ancestor being either A or B (for which a formal procedure is developed in Maslova 2003).

2) After reconstructing the ancestors, look for changes. Ask what is the context of change: correlated or uncorrelated.

3) Count the changes weighting them by the inverse of the branch length and rate on which they occurred. Joint changes on a small branch count more than on a longer branch. Joint changes involving slower traits count more than those involving faster
traits. For directional traits, count the appropriate rate for the direction the change was observed in. Counts of separate changes of the traits, but within a lineage if they occur close to each other, should be counted as correlated rather than uncorrelated.

Hence, it seems unlikely that all word order universals rejected by Dunn et al. will turn out to be Type II errors in Dunn et al.’s analysis. It may be that problems with other assumptions may render some of these rejections questionable (see §4.3-§4.4); but without further testing, one cannot be certain. We will return to this issue in the conclusion.

4.3. Assumptions (1g) and (1h): Language contact and constancy

The next assumption by Dunn et al. is that word orders are co-inherited with the lexical cognates that determine the language phylogeny in their model. Put in linguistic terms, that is the assumption that word order change (or stasis) is largely an internal phenomenon, that is, that it does not typically come about via language contact.

Empirically, this is a quite problematic assumption. Much word order change arises via language contact. A clear example of this is the shift to OV and postpositional order in a small number of Austronesian languages in Dunn et al.’s sample. These languages are in close contact with non-Austronesian languages in Papua New Guinea, which are overwhelmingly OV and postpositional. Similarly, the Indo-Aryan languages are likely to be OV and postpositional due to their contact with Dravidian languages when the ancestral Indo-Aryan speakers migrated into South Asia. It is also likely that word order
stasis is the result of contact rather than retention through the lineage. It is difficult to disentangle the contribution of contact from common ancestry, because most of a language’s neighbors are also its phylogenetic sisters or cousins. Jaeger et al. (this issue), examining the phonological metric used in Atkinson (2011), found that there were detectable effects of both phylogeny and geography in determining the phonological traits of a language. As a matter of fact, the geographical effects reported by Jaeger et al. are comparable in size to phylogenetic effects.

How likely is the acceptance of contact-induced changes to affect Dunn et al.’s results? Currie et al. (2010) argue that only high levels of contact-induced changes, which they call “horizontal transmission”, would affect the detection of a correlation of traits (or absence thereof) in a phylogenetic trait evolution model. Currie et al. treat “horizontal transmission” in the body of their article as a kind of noise that might interfere with a signal of a trait correlation (positive or negative) that is being propagated by “vertical transmission” (i.e. through a phylogeny). This assumes that the correlation is primarily manifested in “vertical transmission”. (In the discussion section of their article, they suggest that linkage of traits or absence thereof in “horizontal transmission” is also evidence of a causal link or lack thereof between the traits; but they do not provide an integrated model of descent and contact.) Currie et al. also compare their phylogeny-sensitive analysis to a benchmark regression analysis that does not take the possibility of dependencies between the data points into consideration at all, whether by common descent or by contact.

But the alternative assumption that typologists entertain is not one in which contact-induced change is noise introduced into an essentially “vertical” model of language
change, nor one that ignores historical dependence of traits (see §2). It is that the geographical traits are inherited (that is, form lineages) of their own. That is, there is a lineage of descent for geographical traits, but it is different from (or at least, not necessarily the same as) the phylogeny defined by the cognate sets. The terms “vertical” and “horizontal” transmission are rather misleading. In both cases, transmission is “vertical” (i.e., forms a lineage). It so happens that a large amount of basic vocabulary is typically co-inherited as a bundle; this is what is called “vertical” transmission, and any lineage that does not bundle with the basic cognate lineages is called “horizontal” transmission. Statistically speaking, this results in a different correlation structure of the “horizontally” transferred traits than assumed in Currie’s analysis.

So it does not appear that Currie et al. (2010) satisfactorily answers our second question. In fact, we conclude that allowing for contact-induced change would make a substantial difference in the results. The phylogeny based on cognates does not accurately reflect the actual histories (lineages) of the word orders/word order changes in the synchronic distribution. Calculating whether or not word order changes are correlated based on the cognate phylogeny is calculating the correlations on the wrong lineages, to a great extent. (For example, some changes that are treated as distinct because they are on separate phylogenetic branches, such as the word order changes in Cora and Tepehuan in the Uto-Aztecan tree in Figure 1 of Dunn et al. 2010, may belong to single areal lineage [see Dryer, this issue].) It is possible, indeed likely, that some word order lineages are co-inherited with the cognates (what historical linguists traditionally call internal change), even if many lineages are not (external change).
Another problem that arises if traits evolve both along the cognate phylogeny lineages and along other lineages (i.e. via contact), is that the causal model and/or the rates of change for internal change and external change are different, and indeed, language contact is so varied that there is probably no one causal model, let alone rate of change, for different types of language contact. But the BayesTraits model needs to assume some sort of constancy in the trait evolution process, and combining two (or more) distinct processes may run into problems with this. A simple constancy model would be one in which rates of change are fixed. However, BayesTraits Discrete (the module of BayesTraits used by Dunn et al.) also allows the option of implementing a so-called covarion model of trait evolution. The covarion model allows for traits to vary their rate of evolution within and between branches, “absorbing” otherwise unmodeled evolution in such a rate variation.

BayesTraits implements the simplest covarion model. In this covarion model, there is a single matrix of rate parameters which governs the relative rates of change of the features in question (in this case, word orders). Then there is a “switch” variable, which turns on or off along lineages in the tree, but this switching on/off has a constant probabilistic description that itself will be estimated from the data. Only when the switch is “on” are transitions under the rate matrix allowed. The addition of this random switch variable allows the absolute rate of events relative to the tree branch lengths to vary, but it still fixes the relative rates among the different kinds of events (e.g., ratio of VO->OV and the reverse OV->VO will stay the same, but since the change is allowed only when the covarion is on, the overall rate drops because of the fraction of time the covarion is off, which, of course is random and different in different branches). The covarion model
allows for variation in the rates of change along different branches, but it replaces uniformity in rates of change with another uniformity: the covarion on/off switching probabilities that ultimately control the rate variation are held constant across the tree. In other words, the rate is allowed to vary, but the variance/skewness etc. of the rate in different subtrees are held similar allowing their statistical estimation.

We do not know if the data is sufficient to constrain the extra parameters needed in a covarion analysis without unduly increasing the Type II error rate. In fact, Dunn et al. did not use the covarion model in BayesTraits Discrete for their analysis (Russell Gray, pers. comm.). Contacts, like the one in Papua New Guinea that we mentioned above, often produce correlated changes between closely related languages, leading to a locally increased rate of change in a model (like the one used by Dunn et al.) that ignores such “horizontal” transmission. Without a covarion to absorb this, such instances would be expected to increase the false positive rate for inference of correlation between traits, and is potentially an important issue.

Turning to the third question, how much would it take to include contact-induced change (and determine how much it would affect Dunn et al.’s results)? It would take quite a bit, both in terms of the model and the data. The model would have to change quite a lot because contact-driven change does not follow a tree-like pattern: the typological traits themselves are not usually co-inherited as a bundle, and there is likely to be reticulation. One cannot simply change the inputs and parameters to BayesTraits and rerun it. Empirically, determining the contact lineages is probably an even harder problem than coming up with deeper phylogenies based on cognate sets and sound correspondences. We can use geographical proximity as a stand-in, as Jaeger et al. (this
issue) do for the data in Atkinson (2011). Again, this would be quite a different model from BayesTraits. We are not saying a combined cognate-phylogeny-cum-contact model cannot be constructed for typological trait evolution. But such a model would be quite different from the model that Dunn et al. have used.

4.4. Assumption (1i): Polygenesis

Dunn et al. treat the four phylogenies as independent in arriving at their conclusion that the behavior of word-order changes is different in each phylogeny (and hence lineage-specific). This is not unlike an assumption of polygenesis for these four families, although Dunn et al. do not make this point explicitly. Consider the alternative, namely monogenesis. Many, perhaps most, though not all, linguists believe in a single origin of language, even if current linguistic knowledge prevents us from being able to construct a detailed monogenetic phylogeny of languages.

If one changes the assumption of polygenesis to a monogenesis assumption, that is, that (at least) the four families examined by Dunn et al. ultimately have a single common ancestor, then BayesTraits cannot be used for the analysis. Since BayesTraits evaluates the presence/absence of correlations between typological traits, it would find a single set of correlations (if any) for that single lineage. In other words, there would be no lineage-specific correlations, and hence Dunn et al.’s results would not hold.

This sensitivity to the assumption of polygenesis is, of course, a property of BayesTraits. It would be possible to construct a model that allows the correlations, not (just) the rates of change, to vary across a phylogeny, and so not have the result be
precluded by the assumption of monogenesis. But this would be a different model from the one that Dunn et al. used; and it would also raise questions about other assumptions besides monogenesis.

5. Conclusion

Dunn et al.’s analysis of word-order correlations carries out a type of analysis that many typologists from Greenberg onwards have advocated: the dynamicization of synchronic typology. Taking a diachronic approach models more directly the causal connections between word orders that are presumed to underlie the synchronic word order universals that have been the subject of so much typological research.

However, a model is only a model. A model requires certain assumptions about the diachronic processes and their relationship to the synchronic distributions which are the input to the model. One must ask: what are the assumptions of the model? Are they empirically problematic? How sensitive are the results to the assumptions? What would it take to change the model to allow for more plausible assumptions?

In this comment, we identified three major assumptions that are empirically problematic, where changing the assumptions is likely to change the results, given the model Dunn et al. used, as well as several other assumptions which, while empirically problematic, are less likely to change the overall results. The first major assumption is the absence of assessing the rate of Type II errors (false negatives). The data clearly indicates this is a serious issue. Unfortunately, resolving the issue may not be only a question of carrying out the appropriate computationally expensive power studies on the model. It is
also likely that the shallowness of widely-accepted phylogenies in linguistics means that word order changes are too sparse in the data to merit firm conclusions about Dunn et al.’s hypotheses.

The second major assumption is that typological traits are co-inherited with the cognate sets that define the language family trees (phylogenies). In fact, many typological traits have different lineages, as they are transmitted via language contact. So the trees defined by cognate sets will not reflect that actual lineages of the typological traits, and so any trait evolution model that depends exclusively on the trees will not be an accurate model of word order changes. However, constructing a model that incorporates the effects of contact would lead to quite a different model than the one that Dunn et al. used. Barring the development of such a model, in the foreseeable future, one would need to use external linguistic knowledge to curate the available data, and modify the algorithm to ignore cases of possible contact.

The third major assumption is, effectively, polygenesis. The polygenesis assumption allows Dunn et al. to analyze correlations independently in each of the four families they examine, and thereby allows for correlations to be lineage-specific. If monogenesis is assumed, then BayesTraits could not be used to analyze the data. Dunn et al. would need to test a model allowing variations of correlations against their current model of lineage independent correlations across the entire Proto-World tree including the four families. This is, effectively, a change in the prior used in their analysis. This change will affect their results if the assumed depth of Proto-World is not sufficient to erase the memory of the original word orders.

Adding further tests and improvements to the model of diachronic processes would
allow a linguist to consider the results with more confidence as to their reliability. The results of such a revised analysis may be quite different from those presented by Dunn et al. Whatever those results are, they must be taken in conjunction with the synchronic facts across the large sample of languages available to typologists. For example, consider the remark in §4.2 that the non-correlation of adposition-noun order and verb-object order in Uto-Aztecan is unlikely to be due to a Type II error (false negative); let us leave aside for now that the other major issues we have raised may also affect this result. This would make it appear that the adposition-noun/verb-object correlation is lineage-specific. Yet the synchronic biconditional correlation between adposition-noun order and verb-object order is very strong: a raw count of languages in the April 2011 version of WALS indicates that 94% of the languages conform to this correlation. Hence, if the correlation were lineage-specific, the lineages of the world’s languages would somehow almost all happen to end up conforming to the correlation. This fact would still need explaining (if this is not seen as evidence of retention from proto-World).

Typologists may conclude from this commentary that Greenbergian universals have not been eliminated. But typologists should not throw out the baby with the bathwater: the methods that Dunn et al. have pioneered here, when accompanied by power studies and further developed to incorporate more plausible assumptions about language change, are invaluable tools for diachronic typological research.

Correspondence address:

MSC03 2130, Linguistics
Acknowledgements:

We thank Laura Fortunato, Ian Maddieson and Jon Wilkins for their comments on earlier drafts, and Russell Gray for providing additional details about the model used in Dunn et al. None of them are responsible for the content of this commentary.

References


Figure 1. Diagram for a one trait, continuous time Markov model, which has two rates of change.
Figure 2. Diagram for a two-trait continuous time Markov model, which captures correlations between the evolution of traits A and B. Red arrows indicate rates which correspond to a change of A from 0 to 1, which would be equal in a model of independent evolution (Figure 3).
Figure 3. Model of the evolution of two traits which evolve independently.