Mixed effect models for genetic and areal dependencies in linguistic typology

Mixed effect models for genetic and areal dependencies in linguistic typology

T. FLORIAN JAEGER, PETER GRAFF, WILLIAM CROFT, and DANIEL PONTILLO

1. Introduction

1.1. *Summary of Atkinson 2011*

Atkinson (2011) sets out to test the so-called "serial founder model" against crosslinguistic data on phonological diversity. In his words (Atkinson 2011: Supporting Online Material: 3), the serial founder model predicts that

> [...] during population expansion, small founder groups are expected to carry less phonemic diversity than their larger parent populations. A series of founder events should produce a gradient of decreasing phonemic diversity with increasing distance from the origin.

To test this hypothesis, Atkinson employs a sample of 504 non-extinct languages from *WALS* (Haspelmath et al. (eds.) 2008), for which the number of vowels, the number of consonants, and the number of tones in the language are annotated (Maddieson 2008a, b, c). For the main analysis, these three measures were standardized (i.e., the mean was subtracted from each value, which was then divided by the standard deviation of the measure) and averaged into one combined measure of the total phonological diversity of a language. This normalized phonological diversity measure ranges from -1.19 to 1.68 (mean $= 0.02$). Each language is also annotated for its coordinates on the globe as well as it population size (the number of speakers). The main text of Atkinson 2011 presents the results of a linear regression analysis of normalized phonological diversity against the distance from the hypothesized "origin of language" while controlling for log-transformed population size and its interaction with the distance from the origin (population size data was taken from Gordon & Grimes (eds.) 2005). The origin of language is determined by comparing the model fit for all 2,560 language coordinates found in the version of *WALS* employed by

Linguistic Typology 15 (2011), 281–320
DOI 10.1515/LITY.2011.021
1430–0532/2011/015-0281

Atkinson (*WALS* has since then been updated, see below). That is, Atkinson fits his model 2,560-times, each time assuming a different origin of language. In order to reflect likely migration routes, distances from the respective origin are based on the Haversine distance between points on a sphere (Sinnott 1984) and the requirement to pass through the geographically motivated intercontinental way points summarized in Atkinson (2011: Supporting Online Material: Table S4, referring to von Cramon-Taubadel & Lycett 2008). The best fitting model is found in West Africa. Moreover, the quality of fit decreases with increasing migration distance from West Africa (see also Figure 9 below).

Atkinson's article has received considerable public attention and sparked lively discussion among typologists. In this commentary, we focus on potential issues with the STATISTICAL procedures employed in the paper. In particular, we investigate to what extent the results are robust once genealogical and geographic relations between languages are taken into account. Such concerns about violations of independence due to the failure to account for relatedness between languages play a central role in quantitative research on typology (e.g., Bell 1978, Dryer 1989, Perkins 1989). We show that the statistical approach taken by Atkinson, linear mixed effect regression, provides a powerful way to control for both genealogical and areal dependencies between languages that has advantages over previous proposals, such as separate regressions by language family or by continent or limiting oneself to stratified samples. While Atkinson (2011) includes only controls for genetic dependencies in his model, we introduce two simple ways to extend mixed effect models to account for effects of language contact ("areal dependencies"). These approaches also provide an alternative way to account for genetic relations about which there is high uncertainty.

To ensure comparability between Atkinson's and our analyses, we rely on the metric of phonological diversity employed by Atkinson. For the same reason, we use the same population size estimates and distance estimates employed by Atkinson. This does not mean that we necessarily endorse Atkinson's decisions to use these metrics, which seem to come with serious problems (see Cysouw et al. 2011, Maddieson et al. 2011). Rather, the primary goal of our paper is to provide readers unfamiliar with mixed effect approach taken by Atkinson with an introduction to this powerful statistical approach.

While we find that Atkinson's results replicate after genetic dependencies and language contact are taken into account, we also find two serious problems with Atkinson's analysis. This leads us to ultimately conclude that the results provided in Atkinson 2011 do NOT provide strong support for the serial founder model. The most serious of these problems is the failure to assess the Type I error rate of his approach (i.e., the rate of false rejections of the null hypothesis). In simulations, we find that the actual Type I error rate of Atkinson's analysis is much higher than the conventionally accepted rate (any statistical analysis

has a Type I error larger than 0). We begin with an overview of the statistical issues we address in this commentary.

1.2. *Overview of the issues addressed*

The ordinary linear model fit by Atkinson provides a decent fit against the data ($R^2 = .31$) and the distance from the origin has a highly significant effect in the expected direction: the phonological diversity of languages seems to decline with increasing distance to the language origin ($\beta = -.00004$, $t = 10.9$, $p < .0001$). This effect is illustrated in Figure 1, which shows both the best linear fit and a local smoother that does not assume linearity between the distance from the origin and normalized phonological diversity. The local smoother was added here to provide an accessible visualization of the rather limited non-linearity in the relation between distance from the origin and phonological diversity (which is good, as we will see below).

The reliability of results obtained from a model depends on to what extent the assumptions of the model are met. Fitting a linear model, such as the one above, assumes normality, homoscedasticity, linearity, and that the observations were sampled independent of each other. It also assumes that overly influential outliers have been removed and that multicollinearity is not an issue. For now, we focus on the assumption of independence and return to the remaining assumptions below, where we also explain what they mean. The assumption

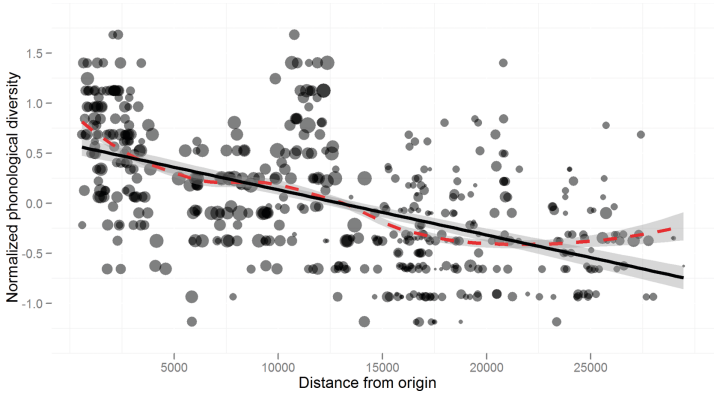


Figure 1. *Normalized phonological diversity plotted against distance from the single-origin with the best fit. Circles represent languages. The size of the circle reflects the number of speakers of that language (as reported in Gordon & Grimes (eds.) 2005). The solid black line shows the best linear fit through the data. The dashed curve shows a non-linear fit by a local smoother (Loess) across all language families. Shaded areas around the two fits indicate 95 % confidence intervals.*

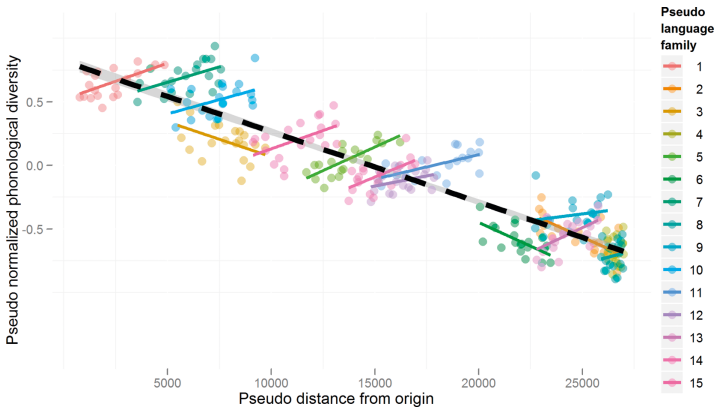


Figure 2. *Illustration of Simpson's paradox based on* SIMULATED *data. Here, a failure to take into account the grouping structure of the data (pseudo language families, indicated by different colors) would result in the conclusion that the (pseudo) normalized phonological diversity (y-axis) decreases with the distance to the language origin (x-axis), although the* OPPOSITE *trend holds for most pseudo language families.*

of independence is obviously violated since the sample employed by Atkinson contains languages that are genetically related, and hence not independent of each other. Additionally, languages may share properties due to extended language contact, leading to another violation of the assumption of independence. This is problematic because languages that are genetically or geographically related, and hence are likely to share certain properties, form so-called clusters in the data (comparable to repeated measures on the same participant in a psycholinguistic experiment). If unaccounted for, such violations of independence are anti-conservative and hence lead to an increased Type I error rate (i.e., higher than intended rates of false rejections of the null hypothesis). Put differently, clustered data can lead to spurious significant effects in the analysis.

A second, intimately related, issue is the possibility of Simpson's Paradox if theoretically motivated grouping structure is not accounted for: it is possible that a trend that is observed across all data points when grouping structure (such as language family) is not taken into account, does not hold within any of the groups or even holds in the opposite direction within groups (Simpson 1951 building on Pearson et al. 1899; Yule 1903). Figure 2 illustrates Simpson's paradox. As much as Simpson's paradox can be a concern, it is crucial to note that fluctuation in the within-group slopes, as observed in Figure 2, does not necessarily mean that there is no effect. We will return to this issue once we have established the necessary concepts and terminology.

To test whether the distance from the origin effect holds once genetic grouping structure is accounted for, Atkinson presents several auxiliary analyses in the supplementary materials of his article. He presents both ordinary linear regressions over family-level data (data that is aggregated by language family) and linear mixed model analyses with crossed random intercepts by language family, subfamily, and genus. Here, we focus our discussion on linear mixed models since we take them to hold considerable promise as a statistical tool for quantitative typology. Linear mixed models are a type of Generalized Linear Mixed Model (Breslow & Clayton 1993), which provide parsimonious ways to account for group level structure in the data while simultaneous assessing effects within and across groups (for additional introductions to mixed models directed at language researchers, see Baayen et al. 2008, Jaeger 2008, Johnson 2009; for additional applications of mixed models to typological data see Cysouw 2010, Cysouw et al. 2011). Atkinson reports that the distance effect remains significant in the predicted direction in all analyses.

In the remainder of this commentary, we will discuss what this does and does not mean. In particular, we show that Atkinson's approach addresses concerns about violations of independence due to genetic relations between language TO A CERTAIN EXTENT. We also show that it is possible to extend Atkinson's analysis to include controls for language contact and that this does not change the results reported by Atkinson. In short, the methods employed by Atkinson are well-suited for typological analysis and have advantages over previous proposals used to account for relations between languages.

However, we also find two serious challenges to Atkinson's conclusion. The first relates to the caveat that Atkinson's model only corrects to a certain extent for violations of independence. This caveat turns out to be a serious one. As we will see, the *WALS* sample employed by Atkinson simply does not contain enough language families with sufficiently many languages to be confident that the claimed distance effect still holds once between-language family variation in the effect are taken into account. The second and, in our view, more serious challenge originates in Atkinson's decision to refit the origin model 2,560 times (in order to find the best origin) to then report the BEST fitting model, where the maximized model fit is a function of the very predictor that Atkinson is interested in (distance to the origin of language). To simplify somewhat for now, this approach has a high chance to find a significant distance effect even if there is none. In other words, the approach taken by Atkinson results in a very large Type I error (see also Cysouw et al. 2011, Jaeger et al. 2011). This is not a principled limitation of mixed models, but rather a problem with Atkinson's use of mixed models (we would like to add, in our view, though that Atkinson deserves credit for pushing the standards of statistical data analysis for typological research; see also Cysouw 2010).

In order to make our assessment of the approach taken by Atkinson accessible to a broader audience of quantitative typologists, we begin by providing a brief introduction to mixed models. We then examine how and to what extent Atkinson's linear mixed model analysis accounts for genealogical effects. The goals of these sections are two-fold. First, we introduce readers unfamiliar with mixed models to this powerful statistical tool. Second, we hope to make the analysis presented in Atkinson 2011 more accessible by replicating it step-by-step. After replicating the mixed model reported in Atkinson 2011, we return to the issue of Simpson's paradox raised above. We discuss what Simpson's paradox does and does not imply and to what extent mixed models can help to address Simpson's paradox. We then present two possible ways to extend Atkinson's model to account for language contact in terms of geographical effects. With the new model in hand, we revisit the search for the best single origin of language under the assumption of a serial founder account. We find that, even after geographical effects are accounted for, the best fit for a single origin model robustly predicts this single origin to lie in West Africa, replicating Atkinson's results. We close with a summary of our analyses and a list of remaining statistical issues, including the large Type I error rate mentioned above.

2. Generalized linear mixed models

Linear mixed models are an extension to the linear model. In the linear model (linear regression), an outcome (or dependent variable), y, is described by means of a linear predictor plus normally distributed noise (often called ε). The linear predictor is a weighted sum of all predictors in the model, so that for each data point i, the outcome y_i is described by (E1):

$$\text{(E1)} \qquad y_i = \beta_0 x_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

The term $\beta_0 x_0$ is often abbreviated as α (the intercept), as x_0 is assumed to be the constant 1 and β_0 refers to the intercept coefficient. The remaining βs are the weights (or coefficients) to the predictors $x_1 \ldots x_k$, such as, in the current case, log-transformed population size, the distances from the language origin, and their product (corresponding to their interaction). The final term, ε, is the randomly distributed noise (writing $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ means that we assume that ε is drawn from a normal, or Gaussian, distribution with mean 0 and variance σ_ε^2). In other words, rather than expecting the outcome to be perfectly described by the linear predictor, we assume that the process that generates the outcome is inherently probabilistic and hence noisy. With this in mind, the ordinary linear regression model presented in Atkinson is described as:

(E2) $$\begin{aligned}&\text{NormalizedPhonologicalDiversity}_i = \\ &\alpha + \\ &\beta_{\text{PopulationSize}} * \log(x_{i,\ \text{PopulationSize}}) + \\ &\beta_{\text{Distance}} * x_{i,\ \text{Distance}} + \\ &\beta_{\text{PopulationSize:Distance}} * \log(x_{i,\ \text{PopulationSize}}) * x_{i,\ \text{Distance}} + \\ &\varepsilon_i, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)\end{aligned}$$

In non-Bayesian data analysis (the current standard in the behavioral sciences), the best coefficient estimates $\hat{\beta}_0 \ldots \hat{\beta}_k$ are typically derived by maximum likelihood estimation.[1] The maximum likelihood estimates of $\beta_0 \ldots \beta_k$ maximize the probability of the observed data given the predictors and the assumptions of the linear model (a normally distributed outcome that is linear in the βs). Statistical inferences can then be drawn over this maximum likelihood model. As a means of illustration, it might be helpful to think of Equation (E2) in geometrical terms. The coefficient estimate, $\hat{\beta}_{\text{Distance}}$, for the distance from the language origin, $x_{i,\ \text{Distance}}$, is an estimate of the SLOPE of the solid black line in Figure 1 once the effects of the other predictors in the models are taken into account.

In the case of Atkinson's study, the hypothesis we would like to test is whether the true slope, β_{Distance}, is smaller than zero (since a negative effect is predicted). This hypothesis is tested based on the estimated coefficient, $\hat{\beta}_{\text{Distance}}$, and its estimated standard error (see below), while controlling for the effects of other predictors, such as population size.

One major shortcoming of the ordinary linear model is that it provides no direct way to account for violations to the assumption of independence. Such violations are bound to arise whenever data points fall into groups (i.e., when subsets of the data are inherently related and hence not independent). Linear mixed models provide an elegant way to account for such grouping structure, thereby re-establishing (conditional) independence. In addition to individual-level noise ε, linear mixed models allow for normally distributed group-level differences centered around the individual level parameters. Atkinson presents a linear mixed model with random intercepts by language family, subfamily, and genus (Atkinson 2011: Supporting Online Material and personal communication), which we can write as:

1. We use the hat notation whenever we are referring to estimates, as opposed to the true underlying – and usually unknown – coefficients.

(E3) $$\begin{aligned}
&\text{NormalizedPhonologicalDiversity}_i = \\
&\alpha_{J,K,M} + \\
&\beta_{\text{PopulationSize}} * log(x_{i,\ \text{PopulationSize}}) + \\
&\beta_{\text{Distance}} * x_{i,\ \text{Distance}} + \\
&\beta_{\text{PopulationSize:Distance}} * \log(x_{i,\ \text{PopulationSize}}) * x_{i,\ \text{Distance}} + \\
&\varepsilon_i, \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\
&\alpha_{J,K,M} = \alpha_0 + a_J + a_K + a_M, \quad a_J \sim \mathcal{N}(0, \sigma_J^2), a_K \sim \mathcal{N}(0, \sigma_K^2), \\
&\qquad\qquad\qquad\qquad a_M \sim \mathcal{N}(0, \sigma_M^2), \\
&\qquad\qquad\qquad\qquad a_J, a_K, a_M \perp \varepsilon
\end{aligned}$$

In this linear mixed model, the intercept is the sum of the ordinary intercept (cf. α in (E2) above) and three adjustments based on the language family, subfamily, and genus (i.e., a_J, a_K, and a_M respectively). Each of these adjustments is assumed to be normally distributed and centered around 0 (additionally, these group-level adjustments are assumed to be orthogonal to the individual level noise ε). These adjustments are called random INTERCEPTS because they adjust the overall intercept α_0 to reflect the – by assumption – randomly distributed group-specific intercepts. In addition to random intercepts, mixed models also allow random SLOPES (i.e., adjustments to the slopes of the predictors, the βs). We will return to this point below.

The model in Equation (E3) can capture genealogical effects on the overall phonological diversity due to three levels of genealogical relations. Remarkably, it does so with only three parameters: the standard deviations of the normally distributed random intercepts (σ_J, σ_K, and σ_M). To illustrate how efficient mixed models are, it is helpful to compare this approach to two common alternative approaches.

First, it is possible to run the ordinary linear regression model shown in (E2) above by group, i.e., separately for each level of a grouping factor. We will call this the by-group approach. For example, we could run separate linear regressions within each language family. There are several problems with the approach. The first problem is that the approach highlights idiosyncrasies at the sacrifice of the bigger picture. Crucially, separate regressions are bound to reveal idiosyncrasies EVEN WHEN THERE ARE NONE IN THE UNDERLYING SYSTEM THAT HAS GENERATED THE OBSERVED DATA. Especially for language families with few languages in the sample that are located in close geographical proximity of each other (and hence not spanning much of a range in terms of the distance from the origin predictor), individual-level noise, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, will create spurious differences in the apparent slope of the distance effect, including apparent reversals of the effect. Although this problem is ameliorated for larger language families, by-family regressions will still find arbitrary differences between language families. The true question of interest is, however, whether there is statistical support for the distance effect once grouping struc-

ture is taken into account. That is, does the distance effect hold globally GIVEN ALL THE IDIOSYNCRATIC DIFFERENCES IN THE SLOPE BETWEEN LANGUAGE FAMILIES? In order to answer this question, the by-group approach requires secondary statistics over the distribution of the coefficients from all the by-group regressions. This is acceptable and, as a matter of fact, has been a standard approach in some fields before mixed models became available (e.g., Lorch & Myers 1990).

The separate regressions derived in the by-group approach are also prone to overfitting. To avoid overfitting a linear regression to the sample, it is generally recommended to have at least 10 to 15 data points per parameter in the model (for references, see Jaeger 2011: 170). Since the model that we wish to test has three parameters ($\beta_{\text{PopulationSize}}$, β_{Distance}, and $\beta_{\text{PopulationSize:Distance}}$), we would be advised to have at least 30 to 45 languages within each language family that we want to include in our test. There are, however, only three language families with that many languages in the sample (Indo-European, Austronesian, and Niger-Congo with 30, 42, and 62 languages, respectively). By far most language families in the sample are represented by only one language each (69 out of the 109 families). If we cannot even account for family relations, this approach is certainly not feasible if we would like to take into account genetic relationships beyond the family level, such as subfamilies and genera.[2] This would be regrettable since we show below that these levels do carry information, in line with typologists' intuition.

A second alternative to mixed models is to expand the ordinary linear regression model in (E2) to contain predictors that distinguish between all LEVELS of the grouping factors. For a grouping factor with k levels (e.g., k different language families), we would require $k-1$ orthogonal contrasts. Since there are 109 language families, 20 subfamilies, and 278 genera in the sample, to capture group-level effects for all of three grouping factors separately, we would have to add $108+19+277=404$ parameters to the linear model (two orders of maginitude more parameters than the mixed model approach requires!). Even if we only distinguish between the 278 genera, this would require 277 additional parameter and hence at least 3,000 data points, which we don't have.

The efficiency of mixed models in accounting for group-level effects is attractive since fewer parameters means that the model is less likely to overfit

2. Admittedly, the alternative approach described allows for the slope of the predictors (e.g., distance from the origin) to vary freely by language family, which the linear mixed model in Equation (E3) does not. The alternative approach described here hence is more comparable to a mixed model with random by-family slopes for the predictors (see below). Such a mixed model still requires considerably less data (since it has fewer parameters). This advantage in terms of the number of parameters comes at the potential cost that the differences in the slopes between language families are assumed to be normally distributed, an assumption that can be tested.

(reducing the chance of spurious effects) and more power to detect real effects. Mixed models hence promise to be a particularly useful tool for quantitative typological research, as typological research frequently faces serious challenges from data sparseness and additional data are difficult or impossible to gather. For a more detailed introduction to linear mixed models for language researchers, we refer to Baayen et al. 2008 (for more advanced introductions, see Bates forthcoming, Gelman & Hill 2007, Pinheiro & Bates 2004). Another potential benefit of generalized linear mixed models is that they are suitable for not only continuous, but also categorical data (e.g., count data), which are common in linguistic analyses (e.g., mixed logit models and mixed Poisson models, see Breslow & Clayton 1993). For an introduction to these models for language researchers, we refer to Jaeger 2008 and Johnson 2009.

Next, we describe the linear mixed model from Equation (E3) in more detail. This is the model that Atkinson refers to in the supplementary materials (pp. 5–6). As in Atkinson 2011, the model was fit in the freely available statistics software R (R Development Team 2010), using the function *lmer* from the library *lme4* (Bates & Maechler 2010). Throughout this paper, we provide references to packages (libraries) for R that might be helpful for typologists interested in employing ordinary or mixed linear models in their research.

3. Accounting for genetic relationships

3.1. *Random effect structure*

An examination of the R^2s associated with the random effects by language family, subfamily, and genus reveals the strength of genealogical effects. The ordinary linear model in Equation (E2), which does not have random effects, accounts for 30.8 % of the overall variance in phonological diversity between languages. After adding all three random intercepts, yielding the model in Equation (E3), the model accounts for 64.6 % of the variance (54.7 % for normalized vowel diversity, 44.3 % for normalized consonant diversity, and 79.1 % for normalized tone diversity).

Conveniently, mixed models make it possible to test to what extent the inclusion of any specific random effect in the model is justified. Thus, we don't have to assume that genetic relationships are best modeled in terms of random intercepts by language family, subfamily, and genus. For each of these three possible genealogical grouping factors, we can ask whether its inclusion in the model is statistically justified.

Here, we ask whether we can remove any of the random intercepts, starting with the one reflecting the smallest groupings (genus). This is achieved by a χ^2-test over the difference in deviance between the model in Equation (E3) against the same model without a random by-genus intercept. This test assesses whether the additional complexity associated with the random by-genus inter-

Table 1. *Coefficient estimates, Standard error estimates, and p-values for the predictors in the mixed model described in Equation (E3). P-values are based on 20,000 MCMC simulations.*

	Coefficient estimate	Standard error	p_{MCMC}	
Intercept	.134	−.084	< .1	+
Population size (logged)	.017	.007	< .03	*
Distance from origin (in 1,000 km)	−.029	.006	< .0001	*
Interaction	−.001	.001	> .7	

cept improves the model quality (its fit against the data) significantly.[3] While by-subfamily and by-family intercepts significantly improve the model's quality of fit ($\chi^2_\Delta s > 9.9$, $ps < .002$), by-genus intercepts do not contribute significantly ($\chi^2_\Delta(1) = 1.7, p > .19$). Since the qualitative results reported below do not depend on the inclusion of the by-genus intercept and since genus is a theoretically motivated grouping factor, we report the results from a model with all three genealogical grouping factors (i.e., the model from Equation (E3)).

The coefficient estimates for the three predictors of interest are given in Table 1, for which *p*-values were derived using Markov chain Monte-Carlo sampling (henceforth MCMC sampling).[4] To make the coefficient estimate for the distance from the language origin easier to interpret, distances were measured in 1,000 kilometers.

Since all predictors were centered (by subtracting their mean from each of their values), the intercept estimate $\hat{\alpha}$ encodes the overall predicted mean normalized phonological diversity. In line with the serial founder account, the model returns a highly significant effect of distance to the origin: with every 1,000 km from the origin the best fit to the data predicts a decrease in the normalized phonological diversity of about .03 points, corresponding to about 1 % of the total range of the normalized phonological diversity of languages in the sample (which, as stated above, ranges from −1.19 to 1.68). This effect is sig-

3. Deviance is a measure of model quality based on the model's log likelihood (to be precise, deviance $= -2 * \log$ (y | model)). For sufficiently large data sets, differences in deviance between two nested models approximately follow a χ^2-distribution with *k* degrees of freedom, where *k* is the difference in the number of parameters between the two models. Two models are nested, if one model contains all the predictors (incl. random effects) of the other model, plus additional predictors. For an introduction and examples, see Baayen et al. 2008 and Jaeger 2008.
4. MCMC sampling is employed here since the Student's *t*-statistic is known to be anti-conservative. The MCMC sampling procedure employed here is implemented in the *languageR* library (Baayen 2010).

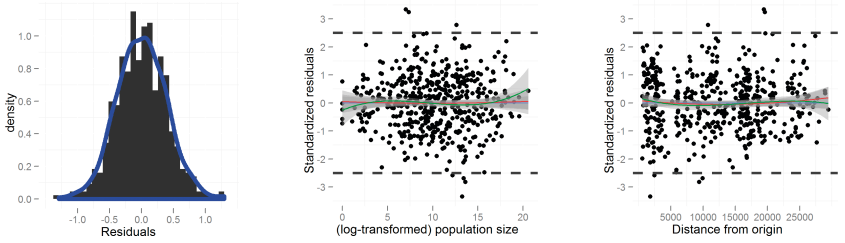
nificant for any combination of the three grouping factors being included as random intercepts (family, subfamily, and genus).

3.2. *Assumptions of linear mixed effect models*

Linear mixed models share several assumptions of ordinary linear models mentioned in the introduction: the outcome is assumed to be normally distributed around a linear predictor, which is assumed to be linear in the coefficients. Furthermore, the errors are assumed not to be correlated with any predictor or with each other (homoscedasticity and no auto-correlation). Violations of these assumptions can lead to unreliable results. A variety of standard tests to assess whether these assumptions are met for a particular data set can be found in the literature on Generalized Linear Models and Generalized Linear Mixed Models (e.g., Agresti 2002, Baayen 2008, Gelman & Hill 2007, Harrell 2001). For the current data set we found that the assumptions of linearity, normality, and homoscedasticity seem to be met or reasonably closely approximated for the current data. Evidence that the assumption of linearity is acceptable for the distance effect comes from the close match between the linear trend and the local smoother in Figure 1. A variety of techniques are available in modern regression programs that allow researchers to relax the assumption of linearity and to systematically investigate non-linear relations within the framework of ordinary and mixed generalized linear models (for introductions see Baayen 2008 and Harrell 2001). Additional tests with so-called restricted cubic splines (Harrell 2001: 20–26) for (log-transformed) population size and the distance effect did not affect any of the conclusions reported here.[5]

Evidence that normality was not violated comes from the observation that residuals were normally distributed (see Figure 3, (A)). Test of the assumption of homoscedasticity returned somewhat more mixed results, although still within acceptable limits. No signs of heteroscedasticity (violations of homoscedasticity) were found for the predictors population size and distance from the origin, which were not correlated with the residuals, as shown in Figure 3, (B) and (C). Further diagnostic plots revealed signs of mild to moderate heteroscedasticity of the residuals by grouping structure, although it is hard to assess the full extent of these violations due to data sparseness (see Appendix B for details). The first caveat to Atkinson's conclusion hence is that the data he

5. Restricted cubic splines and polynomials provide convenient ways to assess degrees of non-linearity in the data. In the statistics software R, the functions *poly()* in the library *stats* (R Development Team, 2010) and *pol()* as well as *rcs()* in the library *Design* (Harrell 2009) interface nicely with procedures used to fit ordinary or mixed regression models (for an introduction, see Baayen 2008). See also the package *gam* for generalized additive models (Hastie 2008).



(A) (B) (C)

Figure 3. *Diagnostic plots for the model described in Equation (E3). The histogram of residuals (the individual-level errors) in (A) suggests normality. Linear (blue), quadratic (red), and cubic fits (green) of log-transformed population size in (B) or distance from the origin in (C) against the standardized residuals reveal no correlations (the shaded 95 % confidence intervals include the zero line at all times). Only six data points fall outside the interval of* -2.5 *to 2.5 standardized residuals (indicated by the dashed lines). Excluding these languages (Austronesian: Iaai, Po-Ai; Niger-Congo: Bisa; Nilo-Saharan: Koyra Chiini; Sino-Tibetan: Garo, Naxi, and Newari) strengthens both the population and the distance effect.*

employed did not contain enough language families with sufficiently many languages to achieve confidence that the assumptions of the linear mixed model are met.

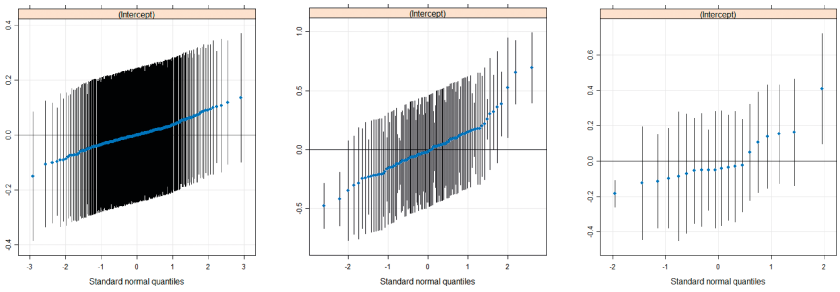
In addition to an evaluation of the assumptions under which a model is fit, it is advisable to test (i) whether overly influential outliers affect the results and (ii) whether multicollinearity affects the interpretation of coefficients or the reliability of standard error estimates. Here, Mandarin is potentially an outlier in terms of its population size, but not an extreme one (z-score $= 2.52$; absolute z-scores larger than 2.5 or 3 are often considered outliers). Indeed, excluding Mandarin from the analysis does not change the results qualitatively. There were no outliers in terms of distance from the origin. Additional analyses removing cases that were outliers in terms of the associated standardized residuals (see, e.g., Figure 3 and Appendix B) did not change the results qualitatively.

Another common issue in any type of regression modeling is multicollinearity. Multicollinearity refers to the presence of high correlations between (sets of) predictors. Multicollinearity can affect the reliability of regression results. Here, multicollinearity was not a concern (fixed effect correlation rs $< .3$). Since none of the models reported in this paper suffered from multicollinearity, we do not report fixed effect correlations below. For methods to detect and avoid issues with multicollinearity, see Baayen 2008 and Jaeger & Kuperman 2009.

The ability of mixed models to efficiently control for shared properties between languages that are members of the same language family, subfamily, or genus is based on the assumption that these differences are normally distributed. This assumption should be assessed when evaluating a model. Figure 4 plots the theoretical vs. actual quantiles of the random intercepts by language family, subfamily, and genus for the mixed model described in Equation (E3). Recall that the only parameters fitted for each random intercept is the standard deviation. It is, however, possible to derive posterior estimates of the random adjustment for each individual group member (e.g., the intercept adjustment for each language family). This is called the Best Linear Unbiased Predictor (BLUP). BLUPs are the modes (the points of with the highest probability) of the posterior distribution of group member adjustments given the model, its parameter estimates, and the data X (including both the predictors $x_1, \ldots, x_k$ and the group membership indicators *J, K*, and *M*). So, for example, for the language family adjustments $a_J, \hat{P}(a_J \mid X, \hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}_\varepsilon, \hat{\sigma}_J, \hat{\sigma}_K, \hat{\sigma}_M)$.[6] In Figure 4, each BLUP (represented by a blue point) is surrounded by its 95 % highest posterior density interval, reflecting the uncertainty about the BLUP, reflected in its distribution $\hat{P}(a_J \mid X, \hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}_\varepsilon, \hat{\sigma}_J, \hat{\sigma}_K, \hat{\sigma}_M)$. If the random differences are indeed normally distributed, the BLUPs should approximately fall on a line between standard normal quantiles > -2 and < 2 (i.e., it should be possible to fit a line in such a way that it touches every interval between -2 to 2). Here, there is no striking evidence for deviation from normality and we have no reason to assume that deviations from normality cause the model to miss important generalizations with regard to language family, subfamily, and genus.

Now that we have ascertained that the normality assumption for the random effects seems acceptable, we can examine the BLUPs that can be derived from the model. As an example, consider Burmese, which is classified as Sino-Tibetan language in the subfamily Tibeto-Burman and the genus Burmese-Lolo. The BLUPs for language family, subfamily, and genus are .385, $-$.097, and .135, respectively. The positive value for Sino-Tibetan correctly captures that Sino-Tibetan languages have higher phonological diversity than the average across all languages. The negative value of the second BLUP suggests that Tibeto-Burman has somewhat less phonological diversity than other Sino-Tibetan languages, and so on. Hence, the mean normalized phonological diversity expected for Burmese solely on its genetic relationships would be .557 ($= .385 + -.097 + .135 + .134$, the overall intercept from Table 1). Appendix A

6. In the statistics software R, BLUPs for mixed models fit with the function *lmer* can be obtained via the command *ranef(model)*. Convenient visualization as in Figure 4 is possible with *dotplot(ranef(model), postVar=T)* and *qqmath(ranef(model), postVar=T)*. All functions are part of the package *lme4* (Bates & Maechler 2010).



(A) (B) (C)

Figure 4. *Quantile to quantile plot of random intercepts by (A) language family, (B) subfamily, and (C) genus in a linear mixed model with the main effects and interaction of (log-transformed) language population and distance from best fit single-origin. Theoretical quantiles – what would be expected under a normal distribution – are shown on the x-axis. The y-axis shows the best linear unbiased predictors (BLUPs) for each level of the random effect. Intervals around dots represent the 95 % highest posterior density intervals.*

provides further information on the BLUPs and how they are related to, but different from group means (the mean normalized phonological diversity of different language families, subfamilies, and genera).

In summary, the current model, which is the same as the one presented in Atkinson 2011, finds that the distance effect holds in the direction predicted by the serial founder account even after adjusting for differences in the phonological diversity between language families, subfamilies, and genera. In the next section, we return to Simpson's paradox introduced above. We ask whether the current model is sufficient to address concerns that the results may be due to Simpson's paradox.

3.3. *Linear mixed models and Simpson's paradox*

Consider the situation in Figure 2, which illustrates Simpson's paradox. Simpson's paradox is particularly relevant to the current discussion, as several researchers have pointed out that, for some language families or regions, the relation between the distance from the origin and phonological diversity seems to go in the opposite of the predicted direction (e.g., Dryer 2011). This is visualized in Figure 5, which shows the distribution of normalized phonological diversity against distance from the origin for the nine largest language families in Atkinson's data.[7] While the linear trend predictor by the serial founder

7. Plots were generated with the libraries *ggplot2* (Wickham 2009), *maps* (Becker et al. 2008), and *lme4* (Bates & Maechler 2010) within the statistics software package R (R Development

effect is observed for Niger-Congo, Nilo-Saharan, and Afro-Asiatic languages (all located on the African continent), a comparably strong opposite trend is observed for Indo-European, Sino-Tibetan, and Austro-Asiatic languages (all located in Eurasia). This could be taken to argue for evidence that the hypothesized effect does not hold across geographic grouping (e.g., continents) or genealogical groupings (e.g., language families).

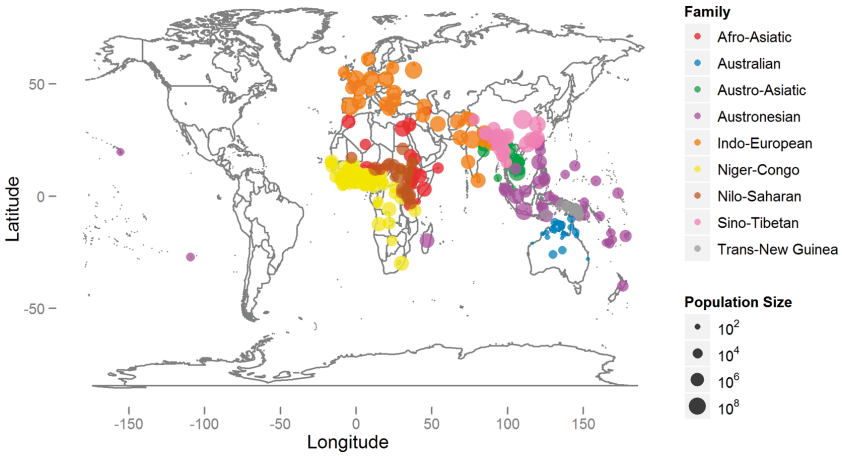
The question we have to ask ourselves is under which circumstances we would want to reject the hypothesis that the distance from the origin predicts phonological diversity. A clear case of Simpson's paradox would be obtained if all within-group trends are the opposite of the between-group trend (see Figure 2 above). Note that even for such a hypothetical extreme, we would have to ask ourselves whether the order of the groups in terms of their distance from the origin is purely co-incidental. For example, if the 109 language families in the sample sort as predicted by the serial founder model and no alternative theory accounts for this order, then this by itself constitutes evidence for the serial founder account. Actually, this is exactly what Atkinson family-level ordinary regression analysis shows (Atkinson 2011: Supporting Online Material: 5–7). Hence, the minimum that any alternative model has to explain is how distance to the origin is either confounded by another variable (cf. Wichmann et al. 2011) or how distance to the origin affects phonological diversity, if not because of a serial founder effect. In other words, Simpson's paradox is less of a problem as it has been made out to be in some of the discussions of Atkinson's article.

In linear mixed models, the question whether there is evidence for a correlation at the family-level is addressed by including random intercepts by language family (and, mutatis mutandis, by subfamily and genus). In other words, we are asking whether there is evidence for a family-level correlation after we have taken into consideration that the different levels of the group (different language families) have different mean phonological diversity and that these differences follow a normal distribution. The results presented in Table 1 suggest that the answer to this question is yes.

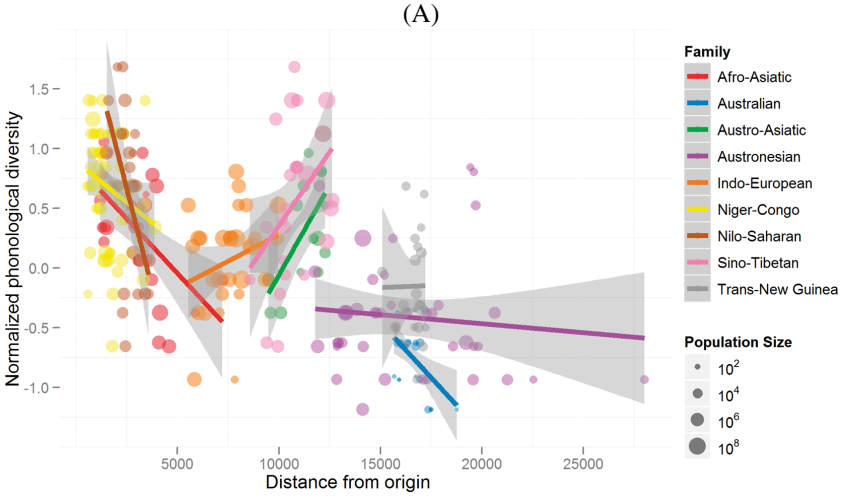
However, strong evidence for a serial founder effect would require that a sizeable portion of the within-group variance in phonological diversity is accounted for by the distance to the origin. Obviously, the clearest case for this hypothesis is obtained if the same trend that is observed between-group also holds within all groups (e.g., within all language families, subfamilies, etc.). This is, however, unlikely to be observed. Any observed data will contain noise (e.g., due to measurement error, misclassification, etc.). Even if the effect we are interested in is large compared to both within- and between-group noise,

---

Team 2010). The code for all plots and analyses is available at http://hlplab.wordpress.com/2011/07/13/glmm-for-typologists/.



(A)



(B)

Figure 5. *(A) Distribution of the nine largest language families in the sample (at least 16 languages each). Circles represent languages. The size of the circle reflects the number of speakers of that language as reported in* WALS. *The color of the circle reflects the language family. (B) Normalized phonological diversity plotted against distance from the origin for the same subset of languages. Solid colored lines show the best fit linear trend with 95 % confidence intervals (shaded area) by language family.*

we might not find the same trend for all groups. Especially, for smaller groups, such as language families that are represented by only a few languages in the sample, we would not expect the trend to hold within each group even if the effect is real.

Within the generalized linear mixed model framework, we can formulate our question as follows: does the predicted effect of distance to the origin hold after we account for difference between language families, subfamilies, and genera in terms of both the overall phonological diversity in that group AND differences in the (slope of the) distance effect (β_{Distance}) within that group? Once we formulate the question like this, it becomes clear that, ideally, we would like to test whether the distance effect holds once we add random by-group slopes for the distance effect. For example, we could add random by-family slopes for the distance effect (β_{Distance}) to the model, thereby allowing the distance effect to vary by language family (cf. Figure 5b). The distribution of between-family differences in the slope β_{Distance} are assumed to be normally distributed – parallel to the assumptions for random intercepts described above.[8] And, just as for the random intercepts, we can employ model comparison to assess whether the random slopes are justified. At least this is possible in theory. In practice, data sparseness can make it difficult to definitively conclude whether random slopes are warranted by the data, as we will see in the next section.

3.4. *Are random slopes for genetic grouping structure required?*

Here, we begin our investigation of random slopes by testing whether random slopes by language family are justified for the distance effect. However, the resulting model does not converge on the full data set. This is due to the large number of small language families in the sample. Out of the 109 language families in the data, 69 are represented by only one language. Only 26 language families are represented by at least four languages. For language families with fewer member languages in the sample, there is simply not enough data to fit both random intercepts and slopes by language family (especially, once we consider that the one language representing a language family is also used to estimate the by-subfamily and by-genus intercepts, as well as the population and distance predictor). At this point, there are two choices: either we get more data

8. Although not technically required, it is recommended (and the default in the mixed model function employed here, *lmer* from the package *lme4*, Bates & Maechler 2010) to include terms for the co-variance between different random effects associated with the same grouping structure. For example, if we add a random by-family slope for the distance effect to the model in Equation (E3), we would by default also add a term for the covariance between the random by-family intercept and the random by-family slope for the distance effect. Below we follow this convention without further discussion (for further detail, see Baayen et al. 2008, Pinheiro & Bates 2004).

or we try to exclude language families with too few languages in the sample. The first approach is more desirable but beyond the scope of this commentary.

The second approach is feasible but results in a catch 22: On the one hand, a model with random by-family slopes for the distance effect will only converge if enough of the language families in the sample contain a sufficiently large number of languages. On the other hand, the sample we analyze still needs to be sufficiently large to be able to find effects. Here, we add random by-family slopes for the population and distance effect as well as the interaction to Atkinson's model and refit it on subsets of the data with only language families with at least k languages in the sample. Whenever the model did not converge, we simplified the random effect structure, first by removing the random by-family slope for the interaction (which always prevented convergence when included), then by removing the random by-family slope for the population effect, and finally by removing the random by-family slope for the distance effect. The same stepwise model simplification procedure was applied when the model converged but model comparison revealed that a random slope was not required.

The result of this process is summarized in Table 2. Models with ANY random by-family slope only converge once only language families with at least four languages are included. Support for Atkinson's conclusion comes from the fact that the distance effect remains significant for this model (columns "4" and "< 7"). It is, however, possible that these subsets of the data still do not contain enough languages per family to find significant random slopes. Interestingly, random by-family slopes ARE justified only when language families with at least seven languages are included. Once random slopes are included in the model, the distance effect becomes insignificant.[9]

What should we make of this? First, we should note that Atkinson's results always replicated for a model with only random intercepts by language, subfamily, and genus: the distance effect remains significant in this model even when only 258 languages, representing eight language families, are left in the sample (not shown in Table 2; in the same model, the population size effect loses significance once only 17 language families are left). However, there IS evidence that random by-family slopes are required. On the one hand, once random slopes are included, the distance effect essentially has become a between-group predictor, dramatically lowering the power to detect an effect. Indeed, power simulations reported in Appendix C suggest that the current data simply do not contain enough language families with sufficiently many languages to detect an effect even when random slopes are included.

9. We note that the distance effect is also insignificant when only the nine largest language families with at least 16 languages are included, which reflects the intuition we arrived at when looking at Figure 5.

Table 2. *Results of linear mixed model with the maximal random effect structure justified for language family based on exclusion of language families with less than k languages in the sample (* = significant; + = marginally significant; empty cells indicate that the random slope or predictor was not significant; n.c. = no convergence, i.e., it was not possible to include this random effect in the model). All models contained random intercepts by language family, subfamily, and genus.*

Minimum number of languages per family	1	2	3	4	<7	7	8	9	<16	16	<20
Remaining families in sample	109	50	31	26	22	17	15	13	10	9	8
Remaining languages in sample	504	445	407	392	376	346	332	316	289	274	258
Random by-family slopes for:											
Population	n.c.	n.c.	n.c.			*			*	n.c.	n.c.
Distance from origin	n.c.	n.c.	n.c.			*	*	*	*	*	*
Interaction	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.
Significant effect of:											
Population	+	+	+	+	+						
Distance from origin	*	*	*	*	*						
Interaction											

In short, the *WALS* data set employed by Atkinson does not provide enough power to detect the distance effect in a model that would be required to answer the question whether the distance effect holds after by-family variation in the slope of this effect is accounted for. Hence, further data will be required to convincingly test the predictions of the serial founder account. This adds a considerable caveat to Atkinson's conclusion.

4. Accounting for language contact

4.1. *Language contact and geography*

One potentially serious issue with the model presented in Atkinson 2011 is the lack of control for language contact. Among typologists, there is rather strong agreement that languages that are spoken in close geographical proximity (in terms of travel distances) of each other over many generations tend to share many features. The assumption is that, on average, geographical proximity tends to correlate with language contact. The Balkan Sprachbund located in Southeast Europe is a classic example: genealogically, the languages of the Balkan Sprachbund fall into five distinct subfamilies of Indo-European (Albanian, Hellenic, Romance, Slavic, and Indo-Aryan). Yet Balkan languages share many grammatical properties. Unfortunately, genetic and geographic groupings are often closely correlated (see also Figure D-1 in Appendix D), so that it can be difficult to disentangle effects of language contact from genealogical ef-

fects. Still, recent studies provide evidence that language contact co-determines typological distributions over and above genealogical relations (Cysouw (in press), Stoneking 2006; see also Croft et al. 2011). Here, we present two ways to account for effects of language contact due to geographical proximity within the mixed model framework.

The first approach is to add to the model random effects for geographical grouping structure, such as continent. While this is a simple and efficient approach, geographical groupings are arguably somewhat more arbitrary than the genealogical groupings. Additionally, treating geographical relations by means of random effects misses the generalization that, for example, continents differ in mutual proximity and accessibility (cf. Africa-Asia vs. Africa-South America). For this reasons, we also describe an alternative way to account for geographical effects: we model the amount of "spillover" of the phonological diversity from neighboring languages as a function of the distance between languages. Ultimately, neither of the two approaches does justice to the complexities involved in modeling the extent of contact between languages over the course of history. The goal here is to explore viable ways to include controls for effects of language contact in typological studies. The emphasis lies on VIABLE, in that there is currently no database that provides a standardized measure of the amount of language contact between languages.

4.2. *Modelling language contact by means of random effects*

To account for language contact by means of random effects, we extracted two geo-cultural grouping factors from *WALS*: continent with six distinct levels and country with 106 distinct levels. Several typologists have proposed continents as a relevant geographical grouping structure (Dryer 1989). Country as a geographical grouping structure, on the other hand, is not generally considered by typologists, presumably since many country boundaries neither reflect obstacles to migration nor cultural divisions. Here we consider country as a grouping structure for three reasons. First, country data was readily available from the *WALS* website, whereas more appropriate regional annotation reflecting cultural ties between adjacent language populations would require costly annotation. Second, in many cases – although admittedly not always – country structure does reflect local groupings. Finally, under the assumption that geographically closer languages have on average more language contact with each other, country structure provides a convenient way to capture a large proportion of the most extensive language contact.

To assess the effect of countries and continents, we added random intercepts for both terms to the model from Equation (E3). We then tested whether removal of a random intercept significantly worsened the model in terms of the χ^2-test over difference in model deviance discussed above. Following con-

ventions for stepwise model comparison, the criterion for excluding a random intercept was set to $p_{\chi^2_\Delta} > .1$; the criterion for including an intercept was set to $p_{\chi^2_\Delta} < .05$. The result of this comparison process is shown in Figure 6. Adding a random by-continent intercept to the model from Equation (E3) (i.e., going from box 4 to box 3 in Figure 6) improves the model only marginally ($\chi^2_\Delta(1) = 2.6, p = .1$). Furthermore, this weak effect is completely subsumed by a random by-country intercept: When the by-continent intercept is removed from a model with all five random intercepts (box 1 to box 2), the resulting simpler model provides just as good a fit against the data ($\chi^2_\Delta(1) = 0.3, p > .5$). However, when the by-country intercept is removed from the model with all five intercepts (box 1 to box 3), this results in a significantly worse model ($\chi^2_\Delta(1) = 39.0, p < .0001$).[10] The model that is most strongly supported by the data contains random intercepts for language family and subfamily as well as for country, but not for continent. Similar to the test presented in the previous section, when only random intercepts by genealogical grouping were considered, the effect of genus does not quite reach significance (e.g., box 2 to green box, $\chi^2_\Delta(1) = 3.2, p < .08$). Language family and subfamily remain significant improvements to the model ($\chi^2_\Delta(1)$s $= 9.4$, ps $< .002$). In short, while continent has been proposed to be an appropriate grouping factor to account for geographical effect, such a grouping is not supported by Atkinson's data.

Table 3 reports the results for the model with the random effect structure best supported by the data (the green box from Figure 6). The inclusion or exclusion of genus does not affect the conclusions reported below (i.e., the results corresponding to box 2 are the same as those corresponding to the green box). The effect of population size is no longer significant ($p_{MCMC} > .16$), but the effect of distance from the origin remains highly significant and the effect still goes in the predicted direction ($p_{MCMC} < .0001$). The interaction of these two predictors remains insignificant.

These results suggest that the effect of distance from the best origin holds up even when geographical effects are controlled for by means of random effects, thereby lending further support to the serial founder model, in line with Atkinson's conclusion.

We close this section by noting that we only considered two geographical grouping factors here, neither of which are arguably ideal. Future work could extend the approach taken here to include random effects that identify areas with high degrees of language contact above the country level, but below the continent level (such as the Balkan area). In the next section, we explore an

10. Additional random SLOPES would be barely justified by continents ($\chi^2_\Delta(2) = 6.0, p = .05$) and not at all by country ($\chi^2_\Delta(2) = .1, p > .9$), compared to the respective intercept-only models. Even if these slopes are included in the models, the results of the model comparisons shown in Figure 6 remain unchanged.

alternative and perhaps more principled way to account for language contact based on geographical distance.

Table 3. *Coefficient estimates, standard error estimates, and p-values for the predictors in the updated mixed model with a random by-country intercept. P-values are based on 20,000 MCMC simulations.*

	Coefficient estimate	Standard error	p_{MCMC}	
Intercept	.103	.081	$> .16$	
Population size (logged)	.008	.008	$> .16$	
Distance from origin (in 1,000 km)	−.035	.007	$< .0001$	*
Interaction	−.001	.001	$> .7$	

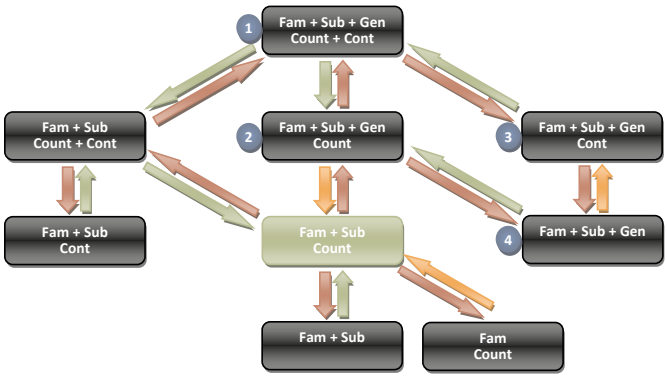


Figure 6. *Schematic illustration of nested model comparison employed to determine the most strongly supported random effect structure. Each box represents a model with random intercepts for the terms in the box (Fam = family; Sub = subfamily; Gen = genus; Cont = continent; Count = country; all models include log-transformed population size, distance from the origin, and their interaction as predictors). Upward arrows represent tests as to whether inclusion of the additional variable(s) present in the upper model is justified compared to the lower model. Downward arrows represent tests as to whether exclusion of these variable(s) from the upper model is justified. The color of arrows indicates whether the corresponding in- or exclusion of a random intercept was justified (green: yes; red: no; orange: marginal). The green model in the middle is the one most clearly supported by the data.*

4.3. *Modelling language contact as a function of migration distance*

The goal of this section is to develop a simple measure of the average phonological diversity of a language's neighbors. This measure can then be included in the model as a predictor to see whether other effects, such as the distance effect, remain significant after the partial variance explicable in terms of areal effects is accounted for.

We follow Atkinson in calculating the migration distance between all pairs of languages based on the Haversine distance between points on a sphere (Sinnott 1984) and the requirement to pass through the way points summarized in Atkinson (2011: Supporting Online Material: Table S4, referring to von Cramon-Taubadel & Lycett 2008). For each language, we then assigned weights to all other languages. These weights were inversely related to the distance to the target language. To be specific, the weight of language j for the calculation of target language k's areal phonological diversity was assumed to decay exponentially with increasing distance to target language k, reflecting a normal distribution centered around the target language and with standard deviation s:

$$w_{k,j} = e^{\frac{-\text{distance}_{k,j}^2}{2s^2}} \tag{E4}$$

The WEIGHTED AREAL NORMALIZED PHONOLOGICAL DIVERSITY of a language k was then calculated by summing over the products of the normalized phonological diversity of all languages j and their weight w_{kj} (the target language's phonological diversity was excluded from this calculation). To put the weighted areal normalized phonological diversity on the same scale as the normalized phonological diversity, the former was normalized by the average sum of all weights:

$$\text{weighted areal normalized phonological diversity}_k = \frac{\Sigma_j w_{k,j} * \text{normalized phonological diversity}_j}{\frac{1}{k}\Sigma_k \Sigma_j w_{k,j}} \tag{E5}$$

With equation (E2) in hand we can compare weighted areal diversity estimates based on different values for the standard deviation s, where higher values for s result in proportionally more weight for closer languages. We compared mixed models as in Equation (E3) updated to also contain the weighed areal normalized phonological diversity in Equation (E5) depending on the value chosen for s in Equation (E4). An iteratively refined grid search returned the best fit for $s = 685$ (see Figure 7). At this value for s, the phonological diversity of languages at a distance of 500 km, 1,000 km, or 2,500 km will be weighted at 77 %, 35 %, and 0.1 % of the weight of a language at a distance of 100 km,

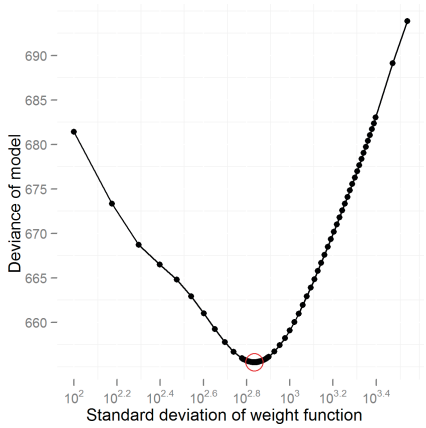


Figure 7. *Quality of fit of a model as in Equation (E3), updated to also include a predictor for the weighted areal normalized phonological diversity (see text). The quality of fit is shown depending on the weight decay rate (the standard deviation, s). Lower deviance indicates a better fit. The best fit was obtained for* $s = 685$ *and is indicated by a red circle. The corresponding model is summarized in Table 4.*

respectively.[11] Figure 8 illustrates the respective weights of languages neighboring Albanian and Hindi.

Interestingly, the best weighted areal normalized phonological diversity ($s = 685$) seems to captures ALL RELEVANT continent-level, subfamily-, and genus-level information as well as some country-level information. If the best predictor for weighted areal normalized phonological diversity is included in a model with random intercepts by language family, subfamily, genus, continent, and country, only the by-country and the by-family intercepts continue to contribute significantly to the model's quality ($\chi^2_\Delta > 16.0, p < .0001$ and $\chi^2_\Delta > 19.0, p < .0001$, respectively). The weighted areal normalized phonolog-

11. The method employed here to find the best value for the s in Equation (E4) has a potential disadvantage: it is quite heavily influenced by the distribution of languages in the sample (i.e., those languages for which information about phonological diversity is available). In other words, the density of languages per area differs for different regions. This means that, for some languages, their closest neighbors IN THE SAMPLE are further away than for other languages. To the extent that this reflects the actual distribution of neighboring languages in the world, there is no problem. However, to the extent that the by-region language density in the sample diverges from the actual by-region language density differently for different regions, this weakens the current approach. For this reason, it is important to note that the results reported below hold for all values of s shown in Figure 7. We also conducted an alternative analysis in which the weighted areal normalized phonological diversity was normalized separately for each language (by dividing by $\sum_j w_{k,j}$ instead of $\frac{1}{k}\sum_k \sum_j w_{k,j}$). This replicates the results reported below.

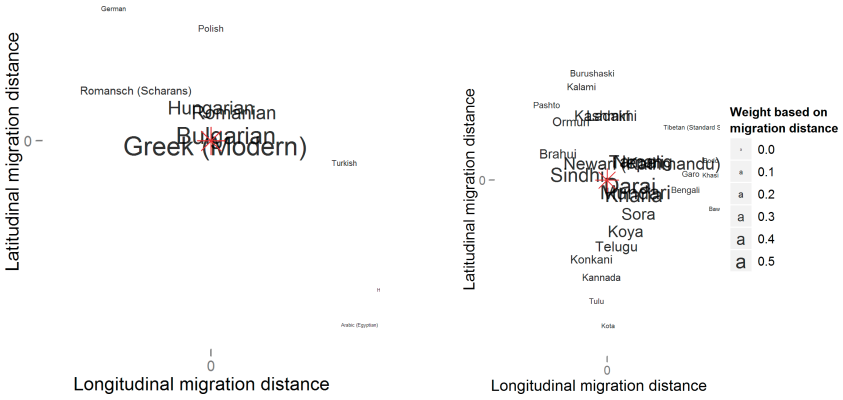


Figure 8. *Illustration of weighted influence of the phonological diversity of neighboring languages on Albanian (A) and Hindi (B). The weighted influence is indicated by font size. Only languages in the sample are plotted. The target languages (Albanian, Hindi) are indicated by a red star. Neighboring languages are plotted based on their longitude and latitude relative to the target language. The relative distances between languages in terms of longitude and latitude were adjusted to reflect migration distances.*

ical diversity remains a significant predictor in all models ($\chi^2_\Delta s > 15.0, ps < .0002$; for further information on the correlation between weighted areal normalized phonological diversity and genetic as well as geographic grouping factors, see Appendix B). We note that the significance of weighted areal normalized phonological diversity is itself not of primary interest here. The critical question is whether the distance effect remains significant after the phonological diversity of surrounding languages is controlled for in the model.

Table 4 summarizes the results of a model with population size, distance from origin, their interaction, and the predictor for the weighted areal normalized phonological diversity. The reported model includes random intercepts by language family and country. The weighted areal normalized phonological diversity had a highly significant effect in the predicted direction: the phonological diversity of languages is to a large extent driven by the phonological diversity that they are surrounded by (the direction of this effect holds for all parameterizations of the standard deviation displayed in Figure 7). Replicating the result reported in the previous section, population size no longer reaches significance, whereas the distance effect remains highly significant in the predicted direction. The interaction of population size and distance from the origin was not significant. As expected, the effect of the phonological diversity of surrounding languages is positive, indicating that languages in close proximity

Table 4. *Coefficient estimates, standard error estimates, and p-values for the listed predictors in a linear mixed model with random by-family and by-country intercepts. P-values are based on 20,000 MCMC simulations.*

	Coefficient estimate	Standard error	p_{MCMC}	
Intercept	.018	.048	$> .3$	
Population size (logged)	.008	.007	$> .14$	+
Distance from origin (in 1,000 km)	−.024	.006	$< .0001$	*
Interaction	−.001	.001	$> .9$	
Weighted areal phonological diversity ($s = 685$)	.390	.078	$< .0001$	*

tend to resemble each other in terms of their phonological diversity. Including random intercepts by subfamily and genus does not change the results qualitatively.

In short, the distance effect based on the best single origin found in Atkinson 2011 continues to remain significant even after various genealogical and geographic effects are taken into account. Of course, the two approaches we have proposed to model effects of language contact are, at best, a reasonable first step. More sophisticated models of language contact could be pursued in future research. First, while the employed migration distances capture some aspect of geography, they fail to account for local terrain and geographic barriers. Second, we might expect that the amount of language contact between two groups of speakers depends on average on the language density of the intervening terrain (i.e., the number of languages spoken in the terrain the speakers would have to cross to be in contact with the other group). Third, language contact does not have to be symmetrical – as a matter of fact, this might be the exception. Ultimately, an adequate investigation of the effect of language contact might require additional databases that capture amount of influence one culture has on another. Alternatively, the extent of lexical borrowing from one language to another could be used to estimate these asymmetric influences. An interim solution could approximate asymmetric effects by entering the relative population sizes into equation (E5) above.

For now, we conclude that, to the extent that we include available controls in the model, the distance result presented in Atkinson still holds. The population effect reported in Atkinson, however, does not reach significance anymore. With what we have learned in mind, we return to Atkinson's search for the origin of language.

5. Re-visiting the search for the origin of language

We repeated the procedure employed in Atkinson to determine the best single origin of language with the updated model we have developed. We extracted all 2,677 language locations recorded in *WALS* (this number is somewhat larger than the one reported in Atkinson 2011, due to recent additions to the *WALS* database; Atkinson, personal communication). Following Atkinson, we then calculated the pair-wise migration distances between all languages (see above). We fit 2,677 separate linear mixed models for each of the possible origins. Each model contained (log-transformed) populations size, the distance to the hypothesized origin (which was different for each model), and the interaction of these two predictors. Additionally, random intercepts by language family, subfamily, and country were included. That is, unlike the model reported in Atkinson 2011, we included a control for language contact in the model. The results reported below are unaffected by the presence or absence of a random intercept for country, random intercepts for all three genealogical grouping factors and/or the predictor for the weighted areal normalized phonological diversity. The results also do not depend on the inclusion of population size in the model.[12]

In all cases, the best origin is predicted to be in West Africa between a longitude of 4.8 to 9.5 and a latitude of -1.25 to 9.33 (incl. Cameroon, Gabon, and São Tomé and Príncipe). Figure 9 shows the model quality for each hypothetical point of origin (assessed as the difference in the deviance between a model with the distance predictor and a model without – both maximum likelihood fitted). The best fits are restricted to Africa. Even the worst fit found for Africa ($\Delta\text{deviance} = 14.5$) is better than the best fit in any other continent (Asia, $\Delta\text{deviance} = 13.8$; cf. the best overall fit, $\Delta\text{deviance} = 24.2$).

In line with the serial founder account, the distance effect has the predicted direction for all models with good fit (i.e., $\hat{\beta}_{\text{Distance}} < 0$, for all the models in Africa and the Middle East), as is illustrated in Figure 10 (Atkinson does not provide the distribution of coefficient values across the 2,560 models he fit, but our replication of his model, the one in Equation (E3), yields a qualitatively similar plot).

12. We note that, for this data set, comparison of ordinary linear models for each hypothetical origin of language happens to return qualitatively equivalent results (the best point of origin differs only minimally and falls within the range of longitude and latitude given above). Of course, the ordinary models each have a considerably lower model quality (e.g., in terms of their R^2) and they over-estimate the significance of the distance effect (as they don't account for the genetic relations between languages).

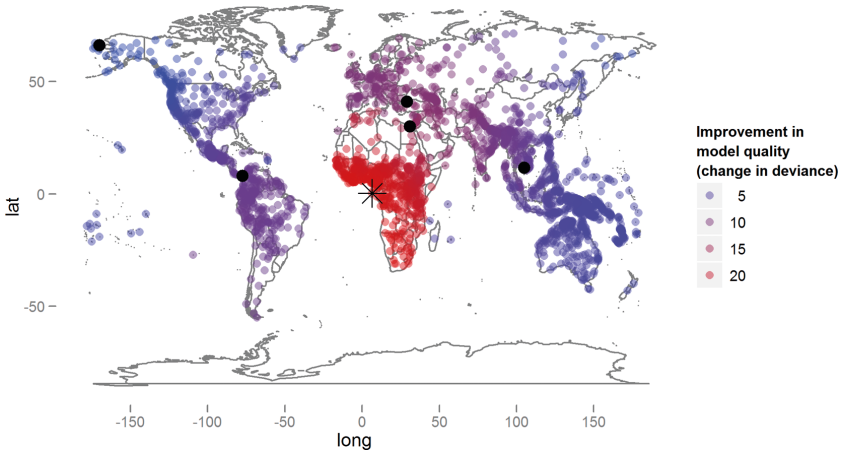


Figure 9. *Model quality depending on hypothesized point of language origin. Better fits correspond to higher improvements in terms of the model's deviance compared to a model without the distance predictor. The best fit is indicated by a star. The solid black dots mark the five way points that inter-continent migration routes are assumed to pass through (see Atkinson 2011: Supporting Online Material: Table S4).*

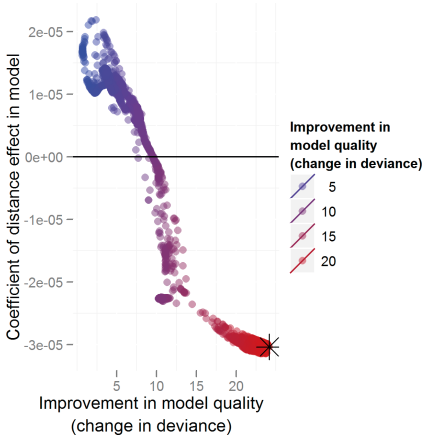


Figure 10. *Coefficient for distance from origin effect ($\hat{\beta}_{Distance}$) depending on model quality (as plotted in Figure 9). The best fit is indicated by a star.*

6. Summary and remaining issues

Atkinson (2011) set out to test the serial founder account, which, he argues, predicts that the diversity of the phonological inventory decreases with increasing distance from the origin of language. To test this prediction, Atkinson employed linear mixed models to account for grouping structure due to genetic relations between the languages in his sample. We have provided an introduction to these models, which we consider to be a useful statistical tool for typological research and we have sketched two approaches to incorporate controls for language contact into such a model, one in terms of additional random effects and one in terms of an aggregate measure to areal phonological diversity. Both approaches revealed large effects of language contact in that language in close proximity resemble each other in terms of their phonological diversity. Interestingly, the second of these approaches turned out to capture lower-level (and hence more local) genetic relations between languages in additional to effects of language contact. This approach may hence provide an efficient way to model genetic relation effects for genetic groupings that are presented by only a few languages in the sample each (e.g., language families with only a few languages in the sample and most, if not all, subfamilies and genera in the sample). While we have focused here on linear mixed models for the analysis of normally distributed outcomes, the generalized linear mixed model framework (Breslow & Clayton 1993) includes models for the analysis of outcomes with different distributions, such as binomially or Poisson distributed outcomes (e.g., count data).

We have scrutinized several of the decisions made in the statistical analyses conducted by Atkinson. In order to maximize comparability of his approach and the tests provided here and in order to achieve the primary goal of this article – to provide an introduction to mixed models for typologists –, we have employed the same measures of phonological diversity, population size, and migration distance (although there are problems with several of these measures; Maddieson et al. 2011; Matthew Dryer, personal communication). We have also abstained from investigating additional hypotheses, such as that the average word length in a language is the primary determinant of its phonological inventory size (Nettle 1998, Wichmann et al. 2011).

Our evaluation of the statistical analyses conducted by Atkinson MOSTLY found his results confirmed for the current sample. We found that even after adding approximate controls for language contact into the linear mixed model, distance from the origin remained a significant predictor of phonological diversity in the expected direction. Model evaluation suggested that most of the assumptions under which the model was fit were met, but that there were potentially some problems with the assumption of homoscedasticity: in a model that has random intercepts for language family, subfamily, and genus, many of

the levels of the grouping factors will be represented by only one language. This makes it difficult to assess whether the residuals of the model exhibit the same variance for all levels of the grouping factors (cf. Appendix B).

However, the sparseness of grouping factor levels with a sufficiently large number of families in the sample also causes a more severe problem: under the arguably more appropriate random effect structure, including random slopes in addition to random intercepts, the distance effect is NOT significant. Atkinson (personal communication) correctly points out that random slopes do not seem to be required for the full data set, but – as we have shown above – this is simply due to the fact that a model with random slopes for the distance effect does not converge on the full data set. If the data is reduced to language families with sufficiently many languages in the sample to successfully fit random slopes for the distance effect, the distance effect becomes insignificant. Further complicating things, our preliminary power simulations indicate that the remaining sample simply does not offer enough power to find the distance effect in a model with random slopes (Appendix C). This leaves us in the most unenviable position that there is evidence that the best model for the complete data set supports Atkinson's claim, while a model that is preferred on theoretical grounds (a model with random slopes) fits only for a small subset, which does not provide enough power to detect an effect (which then, indeed, is not detected).

What DOES follow from this is that (i) there is considerable between-language family variation in the distance effect, but (ii) Atkinson's claim is only supported under the assumption that the languages in the sample that are part of small families are representative for that language family.

This might be acceptable, if it was not for the final remaining issue that we raised in the introduction, but that we have not discussed so far: in order to find the most likely origin of language, Atkinson compares over 2,500 mixed models (one for each language coordinate in *WALS*). The distance effect is then assessed in the model based on the BEST origin. That is, the distance effect is assessed in the model for which adding distance to the model improved the model's deviance the most compared to all other possible origins. This procedure is obviously biased to find an effect for distance to the origin: if there is such an effect for ANY hypothetical origin, it will find it.[13] This is a serious problem with the analysis presented in Atkinson 2011. Admittedly, the DISTRIBUTION of model fits depending on the hypothesized origin seems to support

13. Note that the issue we are raising here is in no way an inherent problem of mixed models, but rather stems from reporting analyses based on models that were pre-selected based on a criterion that refers to the effect of interest (for a discussion of similar biases in the analyses of data from functional magnetic resonance imaging, see Kriegeskorte et al. 2010, Vul et al. 2009).

Atkinson's conclusion (see Figures 9 and 10): all models providing support to Atkinson's conclusion cluster in the same region (Southwest Africa), which also is a likely candidate for the origin of language based on non-linguistic evidence (see references in Atkinson 2011). However, Cysouw et al. (2011) provide evidence that such a geographical clustering of good fits is also obtained by chance under very general assumptions (i.e., even if the serial founder account does not hold; see also Jaeger et al. 2011).

The most appropriate assessment of the error rate would require re-sampling languages in a way that respects their genetic and areal relations as well as their within-family, -subfamily, and -genus distribution of phonological diversity. In other words, what is missing is a simulation that would assess the answer to the question "Given the distribution of phonological diversity within language family, subfamily, and genus, and given the distribution of languages within language family, subfamily, genus, and area (e.g., country), how likely were we to find distributions of model improvements due to the inclusion of a distance to the origin effect that resemble those in Figure 9, except that they have a different origin?"

Here we refer to a modest first step in this direction (for a full report, see Jaeger et al. (2011); see also Cysouw et al. 2011 for similar results): we assessed the Type I error rate associated with Atkinson's analysis by estimating the chance of a significant distance effect based on just (i) the location and genetic relations of the 504 languages in the sample, (ii) the 2,677 possible origins given by the (updated) *WALS* data, and (iii) the distribution of the normalized phonological diversity values in the sample. For each sample of the simulation, the quintuple of language family, subfamily, genus, log-transformed population size, and normalized phonological diversity was randomly re-assigned to the 504 language locations in the sample. We then fit the 2,677 mixed models from Equation (E3) for all possible origins for each of the 10,000 samples. Of the 10,000 samples, model comparison against the baseline model without the distance effect revealed 20.3 % significant improvements. This is considerably higher than the 5% that would be expected under a Type I error rate for $p < .05$ as criterion for significance. Based on the t-value of the models in each sample, we found that for 14.52 % of the 10,000 samples the best model returned a significant main effect of distance to the origin in the expected direction ($t < -1.96$).

In conclusion, while Atkinson is to be applauded for employing a statistical approach that provides a powerful way to control for genetic and areal dependencies in the data, the conclusion that the serial founder model is supported by the sample he analyzes suffers from (i) the inability to control for between-family variation in the effect and (ii) an apparently drastically inflated Type I error rate. Given that others have failed to replicate the effect on alternative data sets (Cysouw et al. 2011), we tentatively conclude that there is, as of now,

no support for the serial founder model from the distribution of phonological diversity across languages.

Received: 14 June 2011
Revised: 6 September 2011

University of Rochester
Massachusetts Institute of Technology
University of New Mexico

Correspondence addresses: (Jaeger, corresponding author) Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14620, U.S.A.; e-mail: fjaeger@bcs.rochester.edu; (Graff) Linguistics, Massachusetts Institute of Technology; (Croft) Linguistics, University of New Mexico; (Pontillo) Brain and Cognitive Science, Computer Science, University of Rochester

Acknowledgements: We are grateful to Quentin Atkinson for generously sharing his data with us and answering our questions about the analyses he conducted. We are grateful to T. Bhattacharya, D. E. Johnson, M. Cysouw, and M. Gillespie for helpful feedback on earlier versions of this commentary, to D. Kleinschmidt for discussions about the modeling of language contact, to T. Stanley for proof-reading, to E. Taliep and X. Wang for extracting the information required for the modeling of language contact from the *WALS* website, and to J. Reyes for help with R and Perl scripting. The work presented here was partially supported by an Alfred P. Sloan Fellowship and a Wilmot Award to TFJ.

Appendix A: More on Best Linear Unbiased Predictors (BLUPs)

In the main text, we introduced the notion of the best linear unbiased predictors (BLUPs). The BLUPs are related to the mean normalized phonological diversity of the respective grouping factors, but they are not the same. Without going into too much detail here, the BLUPs reflect another desirable property of mixed models, called shrinkage (see Gelman & Hill 2007, Kliegl et al. 2010). Shrinkage refers to the fact that BLUPs are shrunk towards the overall mean. The amount of shrinkage is determined by the amount of data available for each level of a random effect and by how far the BLUP estimate is away from the overall mean, thereby avoiding anti-conservativity. This is illustrated in Figure A-1, which plots both the mean normalized phonological diversity for all language families and the corresponding BLUPs. Notice how the BLUP for the language family with the most languages in the sample (Niger-Congo) is identical to its mean, whereas BLUPs for language families with only a few languages in the sample are much closer to the overall mean than a naïve estimate based on the language family's mean normalized phonological diversity would suggest.

In other words, in addition to providing an efficient way to avoid violations of independence, mixed linear models also yield more reliable estimates of group-specific properties (such as the phonological diversity of a language family).

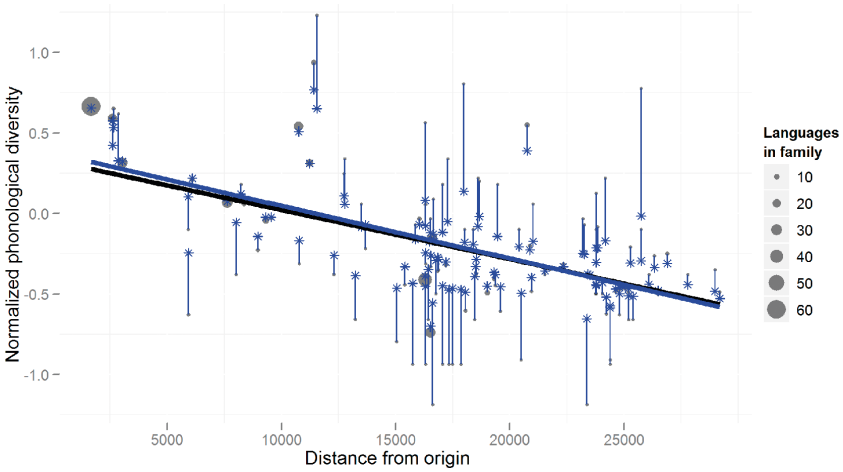


Figure A-1. *Shrinkage illustrated for the random intercepts by language family. Solid circles represent the means across languages in a family (both in terms of phonological diversity and the distance from the origin). Circle size reflects the number of languages in the family. Blue stars indicate the BLUPs for language family intercepts in the Atkinson model with crossed random intercepts by language family, subfamily, and genus. The black and blue solid lines show best linear fits based on the language family means and BLUPs, respectively.*

Appendix B: Diagnostic plots for Atkinson's model

To assess the assumption of normality and homoscedasticity for the individual-level noise (the assumption that the individual-level errors are identically and normally distributed across all levels of the grouping factors), it can be helpful to plot the residuals of the fitted model by group. A full overview of methods to evaluate the validity of mixed models is beyond the scope of this article (for introductions, see Agresti 2002, Baayen 2008, Bates (forthcoming), Jaeger & Kuperman 2009, Pinheiro & Bates 2004). Here, we present two example plots that serve to illustrate that it is difficult to estimate the assumption of homoscedasticity for grouping factors with many group levels that are represented by a small number of individual data points (e.g., language families with few languages in the sample). Figures B-1 and B-2 plot residuals by continent and by combination of continent and language family. Although we did not find continent intercepts to contribute to the model, we include a plot for by-continent variance of the residuals, because continents as grouping factors have received some attention in the typological literature (e.g., Dryer 1989). Overall, the plots suggest normality (for most levels, the residuals are centered around zero). There are, however, signs of heteroscedasticity: the residuals vary

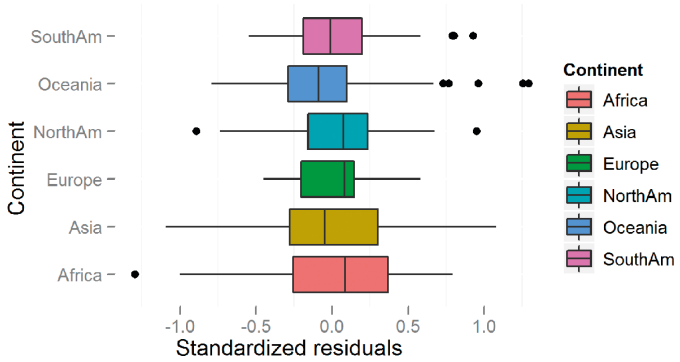


Figure B-1. *Standardized residuals by continent. There are only mild signs of heteroscedasticity (the residuals are not identically distributed between groups). A small number of outliers are also observed (black dots).*

more for some language families than for others (see Figure B-2). Mostly, these differences in variance are relatively small, with the possible exception that the residuals for European languages seem to exhibit less variance.

Appendix C: Power simulations for random slopes

We assessed the power to detect an effect of the distance to the origin in a model with random by-family slopes for the distance effect. Four separate simulations were conducted on subsets of the data that contained only language families with at least 4, 6, 7, or 10 languages. The subsets were chosen based on the results presented in Table 2. For each subset of data, we fitted a linear mixed model with log-transformed population size, distance from the origin and their interaction as predictors, and random by-family intercepts and as well as random by-family slopes for distance. This model was used to estimate the parameters for the simulation (the residual variance, the variances and covariances of the random by-group effects, and the coefficients for the predictors). The results reported below are qualitatively identical if all but the variance and covariance for the random by-family effects are assessed from the model without random by-family slopes. Based on these parameters, 10,000 simulated data sets were created for each subset of data, using that data set's grouping structure (i.e., the 10,000 data sets had the same number of language families as in the actual data and each language family had the same number of languages as in the actual data). These simulations found that the mixed model with random by-family slopes was able to detect the distance effect only in 1 to 4.5 % of all cases.

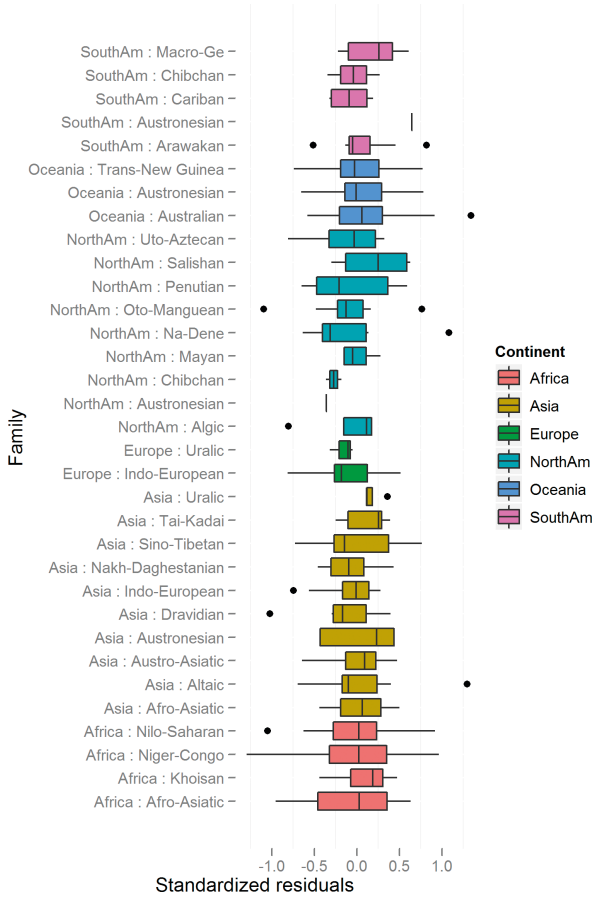


Figure B-2. *Standardized residuals by combination of continent and language family. Only families with at least 4 languages are included. There are signs of heteroscedasticity (the residuals are not identically distributed between groups). A small number of outliers are also observed (black dots).*

This suggests that the current data set does not contain enough language families with a sufficiently large number of languages to assess whether the distance effect holds beyond the estimates between-family variance in the slope of the distance effect. Figure C-1 summarizes the coefficients and *t*-values for the distances effects found in the 10,000 simulation runs.

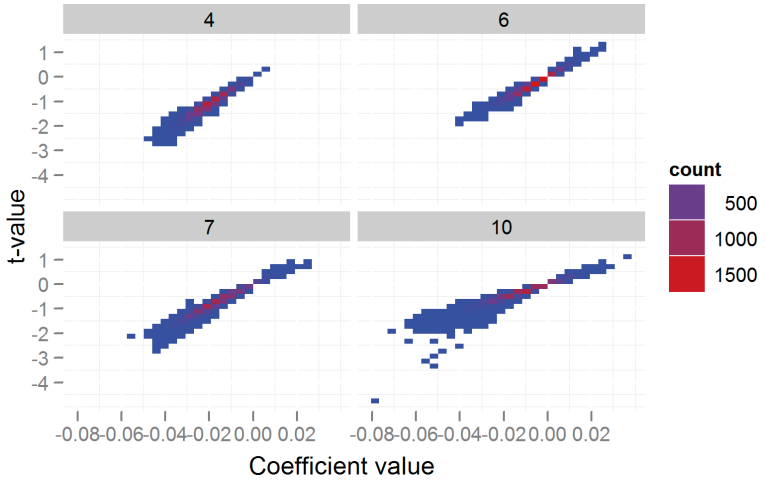


Figure C-1. *Heatmap of the 10,000 simulation runs for each of the simulated subsets of the data. The number at the top of each panel indicates the minimum number of languages per language family in the subset. Simplifying somewhat, a t-value of less than −1.96 would indicate significance (ignoring anti-conservativity of the t-test for the current purpose since conducting 20,000 MCMC simulations on each of the 10,000 runs for the four data sets was not feasible).*

Appendix D: Correlations between genetic and geographic effects

Since genetically related languages often also are located in close geographical proximity to each other, it has been difficult to tease genealogical and geographic effects apart (e.g., Cysouw (in press), Stoneking 2006). Indeed, the correlation matrix in Figure D-1 confirms that genealogical groupings (language family, subfamily, and genus) are moderately to highly correlated with geographic effects. The mixed effects analyses reported in the main text do, however, confirm independent effects of genealogical and geographical effects on normalized phonological diversity.

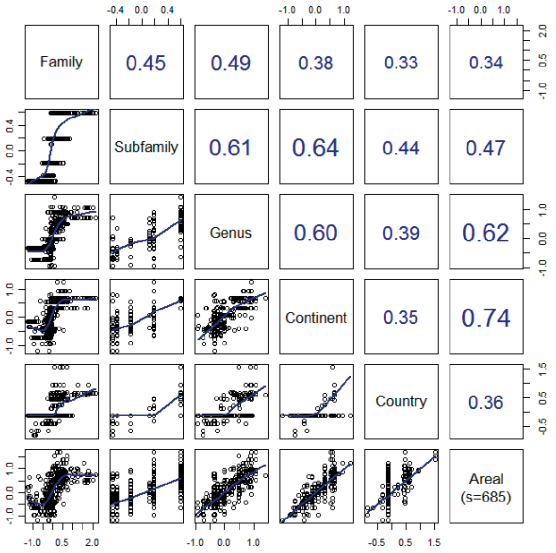


Figure D-1. *Correlation matrix of genealogical and geographic effects on normalized phoneme diversity. For all 504 languages in the sample, the average by-family, by-subfamily, by-genus, by-continent, by-country, and the best by-area normalized phonological diversity was calculated. The upper right part of the matrix shows the Pearson R^2 for each pair of variables. The lower left shows the corresponding scatterplot between the two variables along with a local smoother (blue line).*

References

Agresti, Alan. 2002. *Categorical data analysis*. Hoboken, NJ: Wiley.

Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346–349.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2010. LanguageR: Data sets and functions with 'Analyzing linguistic data: a practical introduction to statistics'. *R package version 1.0*.

Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.

Bates, Douglas (forthcoming). *lme4: Mixed-effects modeling with R*.

Bates, Douglas & Martin Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-37. Retrieved from http://CRAN.R-project.org/package/5lme4

Becker, Richard A., Allan R. Wilks, Ray Brownrigg & Thomas P. Minka. 2008. maps: Draw geographical maps. *R package version 2.0-40*.

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of human language*, Vol. 1: *Method and theory*, 123–156. Stanford, CA: Stanford University Press.

Breslow, Norman E. & David G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88. 9–25.

Croft, William, Tanmoy Bhattacharya, Dave Kleinschmidt, D. Eric Smith & T. Florian Jaeger. 2011. Greenbergian universals, diachrony and statistical analyses. *Linguistic Typology* 15. 433–453.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14. 253–286.

Cysouw, Michael (in press). Disentangling geography from genealogy. In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*. Berlin: de Gruyter.

Cysouw, Michael, Dan Dediu & Steve Moran. 2011. Still no evidence for an ancient language expansion from Africa. Manuscript, submitted.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.

Dryer, Matthew S. 2011. The evidence for word order correlations. *Linguistic Typology* 15. 335–380.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models* (Vol. 3). New York: Cambridge University Press.

Gordon, Raymond G. & Barbara F. Grimes (eds.). 2005. *Ethnologue: Languages of the world*. 15th edn. Dallas: SIL International.

Harrell, Frank E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Harrell, Frank E. 2009. Design: design package. *R package version, 2(0)*.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2008. *The world atlas of language structures online*. 1st online edition. München: Max Planck Digital Library.

Hastie, Trevor. 2008. GAM: Generalized additive models. *R package version, 1*.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59. 434–446.

Jaeger, T. Florian. 2011. Corpus-based research on language production: Information density and reducible subject relatives. In Emily A. Bender & Jennifer E. Arnold (eds.), *Language from a cognitive perspective: Grammar, usage, and processing – Studies in honor of Tom Wasow*, 161–197. Stanford, CA: CSLI Publications.

Jaeger, T. Florian & Victor Kuperman. 2009. Issues and solutions in fitting, evaluating, and interpreting regression models. Paper presented at the 2009 Workshop on Ordinary and Multilevel Models, UC Davis, CA.

Jaeger, T. Florian, Daniel Pontillo & Peter Graff. 2011. Commentary on Atkinson 2011: Excessive Type I error rates. Manuscript, University of Rochester.

Johnson, Daniel E. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed effects variable rule analysis. *Language and Linguistics Compass* 3. 359–383. http://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2008.00108.x/full

Kliegl, Reinhold, Michael E. J. Masson & Eike M. Richter. 2010. A linear mixed model analysis of masked repetition priming. *Visual Cognition* 18. 655–681.

Kriegeskorte, Nikolaus, Martin A. Lindquist, Thomas E. Nichols, Russel A. Poldrack & Edward Vul. 2010. Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism* 30. 1551–1557.

Lorch, Robert F. & Jerome L. Myers. 1990. Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16. 149–157.

Maddieson, Ian. 2008a. Consonant inventories. In Haspelmath et al. (eds.) 2008, Chapter 1. http://2008.wals.info/feature/1

Maddieson, Ian. 2008b. Tone. In Haspelmath et al. (eds.) 2008, Chapter 13. http://2008.wals.info/feature/13

Maddieson, Ian. 2008c. Vowel quality inventories. In Haspelmath et al. (eds.) 2008, Chapter 2. http://2008.wals.info/feature/2