Supporting Online Material for

Still no evidence for an ancient language expansion from Africa

Michael Cysouw, Dan Dediu and Steven Moran

email: cysouw@lmu.de

Table of contents

1. Materials and Methods

1.1. Data

The linguistic parameter that Atkinson (*S1*) investigates is the size of the phoneme inventory of a language. Although the acoustic variation of possible linguistic utterances is basically continuous in nature, humans discretely categorize this continuous variation into distinctive groups, called phonemes. This discretization is language-specific, i.e. different languages have their own structure of distinctive groups. Empirically it turns out that some languages have more groups (i.e they divide phonetic space into more fine-grained distinctive phonemes), while other distinguish less phonemic clusters of sounds.

To investigate variation in phoneme inventory size, it would have been straightforward for Atkinson to use data on the actual number of phonemic distinctions in different languages. Much of what is known about phoneme inventories is based on the *UCLA Phonological Segment Inventory Database* (UPSID; *S2*). The original UPSID sample size of 317 languages was later expanded to 451 (*S3*) and more recently merged with the core language sample for the *World Atlas of Language Structures* (WALS; *S4*). UPSID is publicly available online and is the most widely used data set for investigating issues in phonological universals and typology (*S5,S6*).

Unfortunately, the data as used by Atkinson are only coarse-grained summaries of the slightly expanded version of UPSID as made available in WALS. Although the WALS data includes a few more languages, only a few illustrative aspects of phonemic variation among the world's languages were included, not the complete UPSID data. Specifically, Atkinson only combines the features 'consonant inventories' (WALS 1, *S7*), 'vowel quality inventory' (WALS 2, *S8*) and 'tone' (WALS 13, *S9*) to obtain an estimate of the size of the phoneme inventory. The data as used by Atkinson is thus really only a rough (and as we will show rather biased) estimate of the actual number of phonemes per language. Note that tonal opposition are not included in UPSID, though the data in

WALS is from the same author as UPSID, Ian Maddieson. For our replication to remain compatible with Atkinson's approach, we added tonal marking to the UPSID data.

We will use the UPSID-451 database (*S10*) to illustrate our concerns with Atkinson's approach. There exist better and much more expanded databases on phoneme inventories, but because these are not (yet) publicly available we decided against their inclusion here. Further, we will only use the total number of phonemes as listed for each language in UPSID. It would be much more interesting to investigate the actual variation within the inventories, but such research is too extensive for the scope of this reply (see Section 2.2 below). We removed the languages *Island Carib* and *Lai* from the UPSID data because these languages have not been included in WALS. Further, the language *Ju|hoan* was removed because the inventory size is an extreme outlier (141 phonemes, while mean of UPSID is 29±10.3), prompting discussion about a suitable analysis of its phonemic structure: it might be better to analyze the phonemes of Ju|hoan as clusters of phonemes (*S11*-*S12*).

A central problem with using UPSID in comparison to Atkinson's analysis is that UPSID does not include information about tonal distinctions. We decided to add the WALS data about tone to UPSID to obtain comparable measurements of phonemic inventory size to Atkinson's measurement. Because WALS is not explicit about the exact number of tone distinctions for languages with 'complex tone systems', we approximated the number of tones in such languages with a mean of 4 tone distinctions. Further, because WALS does not provide information about tone for all languages in UPSID, the combination of UPSID plus tone reduces the set of available languages to 411. Finally, in various analyses we will use speaker community size as a factor, but a further 11 UPSID languages do not have any speakers left, reducing the number of usable languages in these analyses to 400.

Finally, note that there are various different versions of WALS available. WALS was originally published as a book in 2005, and we here still use the data from this original version (*S4*). The data was republished online in 2008 with only minimal changes. Atkinson cites this online version, though he added page numbers that refer to the printed original from 2005. Recently, the online version has been renewed to a 2011 version and some new data has been added (*S13*). However, there do not appear to have been any changes in the crucial features discussed in this paper.

1.2. Measuring phoneme inventory size

There are various idiosyncrasies in the WALS data that influence Atkinson's results. First, WALS gives only rough classes of phoneme inventory sizes instead of the actual numbers of phonemes. Second, Atkinson uses consonants, vowels and tone as equally weighted characteristics, while consonants are actually much more frequent than the other kinds of segments; this represents an implicit weighting of specific characteristics of phoneme systems. Third, the WALS count of vowels only includes the number of vowel qualities, ignoring the many other different ways in which vowels are phonemically distinguished in human languages.

The first problem is that the data in WALS only distinguishes approximate classes of phoneme distinctions. For example, for vowel quality inventories only three classes of languages are distinguished, viz. ‘small vowel inventories’ (i.e. languages with 2-4 vowels), ‘average vowel inventories’ (i.e. languages with 5-6 vowels) and ‘large vowel inventories’ (i.e. languages with 7-14 vowels). So, languages with 5 vowels are counted as having more oppositions than languages with 4 vowels, but there is no differentiation between languages with 7 or 14 vowels. Using the actual counts of phoneme oppositions, as available in the UPSID database, is clearly preferable.

The reason that WALS only provides classes of phonemes instead of actual numbers is surely not “due to uncertainty in ascertaining exact inventory counts across languages”, as Atkinson put it (*S1*, p.2). As in every science, there is of course always room left for discussion of individual cases, but the methodology to describe phoneme systems of the world’s languages is well established and clearly sufficiently valid to give accurate estimates of the number of phonemes. The usage of approximate classes in WALS was purely guided by the wish to provide easily accessible maps in the original printed atlas. Distinguishing more than a few classes per map was deemed to be visually displeasing. During the preparation of WALS, the question of the cut-off points for the classes was explicitly discussed, and the author (I. Maddieson) subsequently added an explicit explanation for the definition of the classes to WALS: “the particular cut-off values for the categories were chosen so as to approximate a histogram with a normal distribution, although there are somewhat more languages with inventories smaller than the band defined as “average” than with larger than average inventories” (*S7*).

In practice, Atkinson uses an average of z-scores ($\frac{x-\mu}{\sigma}$) of the numerical values of the WALS classes. This approach is statistically unfounded, because the WALS classes really are on an ordinal scale (all one can say is that languages with ‘small vowel inventories’ have less vowels than those with ‘average vowel inventories’ but not by how much), and not on an interval scale to allow meaningful computation of the mean and standard deviation. It might, therefore, be preferable to use a simple addition of the WALS ordinal levels, although it will be necessary to normalize the number of levels per parameter (WALS 1 distinguishes 5 levels, while WALS 2 and WALS 13 distinguish only 3 levels).

The second problem is immediately obvious when using actual numbers of phonemes instead of the WALS data, namely that almost all languages have many more consonants than vowels. As explicitly noted by Maddieson in WALS, the average number of consonants is much higher than the average number of vowels. The average number of consonants in WALS is minimally below 23 (*S7*), whereas the average number of vowels is almost 6 (*S8*). Yet, in Atkinson's assessment of phoneme inventory size, the vowel inventory size is given equal weight to consonant inventory size, which can be interpreted as an implicit higher weighting of the number of vowels. This problem of implicit weighting is even more severe with tonal oppositions, as this is likewise counted on a par with consonant and vowel inventories. However, the number of tonal oppositions is almost always lower than the number of vowel oppositions. As an estimate of the mean number of tonal oppositions among the world's languages, we will use the following argumentation, based purely on the WALS data as available to Atkinson:

- Languages with 'no tone' are set to having zero tones;
- Languages with 'simple tone systems' are explicitly stated by Maddieson to have only a two-way basic contrast, so we can count them as having two tones;
- Languages with 'complex tone systems' can have a variety of number of tones without concrete specification of the exact number in WALS. We used an approximate average of four tones for these languages.

Given the frequencies of these three types in WALS, the resulting average number of tones in the world's languages is approximately (307·0 + 132·2 + 88·4) / 527 = 1.2. This means that Atkinson's assessments of phoneme inventory size are implicitly strongly biased toward tonal oppositions. Aggravating this implicit weighting is the fact that tonal oppositions show a strong geographic preference for Africa and Southeast Asia, as can be immediately seen in the original WALS map (*S9*). Moreover, if the arguments in (*S14*) are valid, the current geographic distribution of tone is influenced by a genetic bias encoded by two human genes involved in brain size and development, *ASPM* and *Microcephalin*. Importantly, the biasing alleles of these genes most probably postdate the proposed out-of-Africa migration by several tens of thousands of years (*Microcephalin*: 37kya, 95% CI 14-60kya; *ASPM*; 5.8kya, 95% CI 0.5-14.1kya) showing that an important component of the geographic distribution of tone -- and, thus, of Atkinson's assessment of phonemic inventory size -- could very well have no connection to the scenario proposed by Atkinson.

Further, by counting vowel inventory and tonal oppositions as independent characteristics Atkinson introduces yet another implicit weighting, because these two characteristics are actually positively correlated ($r = 0.32$, $p = 0.0015$ using WALS data, with probabilities estimated from a mixed-effects model with genus, family and macroarea as random effects, thus controlling for these types of non-independence between languages). This somewhat surprising correlation is explicitly noted by

Maddieson in WALS (*S9*), and even while it is not clear how exactly this correlation should be interpreted, it results in an even stronger emphasis on languages with tone and large vowel inventories in Atkinson's assessment of phoneme inventory sizes.

The third problem with using the WALS data is that only vowel quality differences are considered in the ‘vowel quality inventory’. There are many more phonetic aspects of vowels that are used by languages in the world to express meaningful differences. Maddieson himself explicitly addresses length, nasalization and diphtongization in WALS (*S8*). Further possibilities, though less frequently attested, are pharyngalization and glottalization. So, Atkinson could, for example, easily have included the WALS feature on vowel nasalization (*S15*) in his phoneme inventory assessment, as this feature is definitionally independent of the three WALS features used. This inclusion might even have been in favor of an African origin, because vowel nasalization is particularly common in West Africa. The UPSID database includes most such vowel oppositions as described for the world’s languages. Note that this aspect argues that there are normally more vowel oppositions than the mean of 6 vowels that WALS 2 indicates.

In summary, Atkinson’s assessment of phoneme inventory size is only a rough approximation of the actual number of phonemes. There are various easy remedies for the most glaring disproportions, like adding a weighting factor to each WALS parameter based on the mean number of oppositions and the number of levels distinguished for each WALS parameter, as shown in (1). This would have been feasible for Atkinson, as it includes only information available in WALS.

(1) Phoneme inventory size = 23/5 · (WALS 1) + 6/3 · (WALS 2) + 1.2/3 · (WALS 13)

As a post-hoc indication of how well these weights fare, we performed a simple linear regression of the UPSID frequencies on the WALS parameters. This results in the following predictive formula in (2), which also shows highest weighting for consonant, and lowest weighting for tone, though the effect for tone is less dramatic than with the formula above. This is probably due to the fact that the assessment of tones in our UPSID data is based on the same WALS data (see previous section).

(2) Phoneme inventory size = 6.5 · (WALS 1) + 4.0 · (WALS 2) + 2.8 · (WALS 13)

To get an impression of how good these approximations are, we correlated them with the actual UPSID counts. Atkinson's average of z-scores reaches $r = 0.604$, while the simple weighted sum in (1) approximates UPSID slightly better with $r = 0.715$, and the post-hoc linear regression in (2) represents the best approximation possible with WALS data, but only reaches $r = 0.719$ (all these correlations are of course highly significant $p < 2.2 \cdot 10^{-16}$). Thus, the simple weighting scheme in (1) is almost the best attainable approximation of UPSID using the WALS data, and is clearly preferable over Atkinson's average of z-scores. Still, all these different WALS-based measures of phoneme inventory size are a rather limited approximation of the UPSID counts.

1.3. Geographic distribution of phoneme inventory size

In most cases, geographic patterns can only be discerned through some kind of geographic interpolation, and Atkinson's global cline is the result of a method of interpolation to be discussed in detail below. However, before trying to induce any global geographic clines, we will first investigate more local patterns of geographic variation. We will show that Atkinson's measurement of phoneme inventory size results in a rather restricted view of world-wide linguistic variation. Additionally, we will show that African languages in the current sample are extremely homogeneous in their inventory sizes. Such homogeneity is rather at odds with any assumed point of origin, as one would have expected large variation instead (as is the case for modern human genetic diversity).

To be able to interpolate geographically, a measure of geographic distance is crucial. To calculate a distance measure between languages, Atkinson uses great circle distance through a few specified waypoints, e.g. to reach America, the distance has to be measured passing through the Bering strait. These waypoints represent an approximation of possible paths of human population movement until a few thousand years ago. However, even if widely used, the main problem with this approach to geographic distance is that the actual number of kilometers between two languages does not seem to be the best approximation to the socio-historical distance between their speakers. Two neighboring languages in Siberia will be measured as being thousands of kilometers apart, while two neighboring languages in Africa are often just a few kilometers apart.

Yet, it is not immediately obvious how to improve on this measure of distance. It is clear that one would like to include climatic, topographic and socio-historical factors in such a measure, but it is difficult to decide what to include and how to obtain the necessary information. We would like to propose a novel approach: instead of conceptualizing the distance between two languages in actual kilometers, we would like to define the distance between two languages as the number of languages that have to be crossed to get from one language to the other. So, the distance between two languages that are 100 kilometers apart might be rather far in areas with high language density, but low in areas with low language density. The central assumption behind this measure is that it is possible to establish the practical impact of external factors (be it climatic, topographic and socio-historical or else) without needing to know which factors really influenced linguistic density and to which extent. The trick is that the current empirically observed language density in the world is a result of any combination of such factors and the actual language density can thus be used as a measure of the factual effect of these unknown factors.

In practice, we removed all sign languages from the 2560 languages in WALS, and we also removed the language *Yazva* because it had exactly the same coordinates as *Komi-Zyrian* (both are Finnic languages). For the remaining 2519 languages, we calculated a Delaunay triangulation between all point-locations for the languages as specified in WALS. The triangulation was not allowed to cross through a few explicitly specified water boundaries (Fig. S1) that humans do not seem to have crossed up until a few thousand years ago. The 2519 languages only represent about one third of the total number of human languages (*S16*), but for the current purpose this sample is sufficient to estimate relative language distances. The distance between two languages is now defined as the shortest path along the graph that results from the triangulation (Fig. S2).

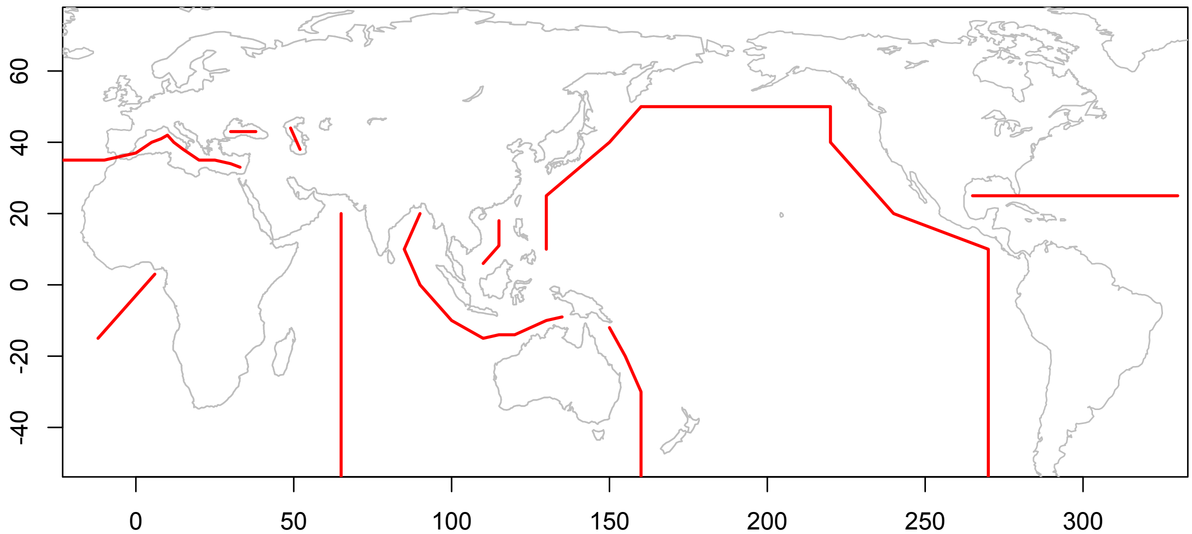


FIG. S1. Hypothesized ancient water boundaries that appear not to have been crossed until a few thousand years ago.

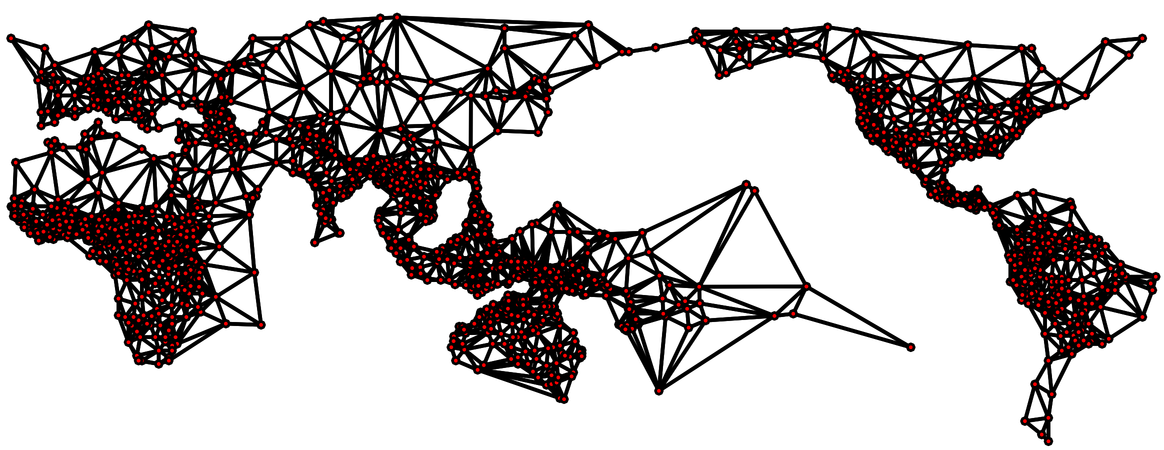
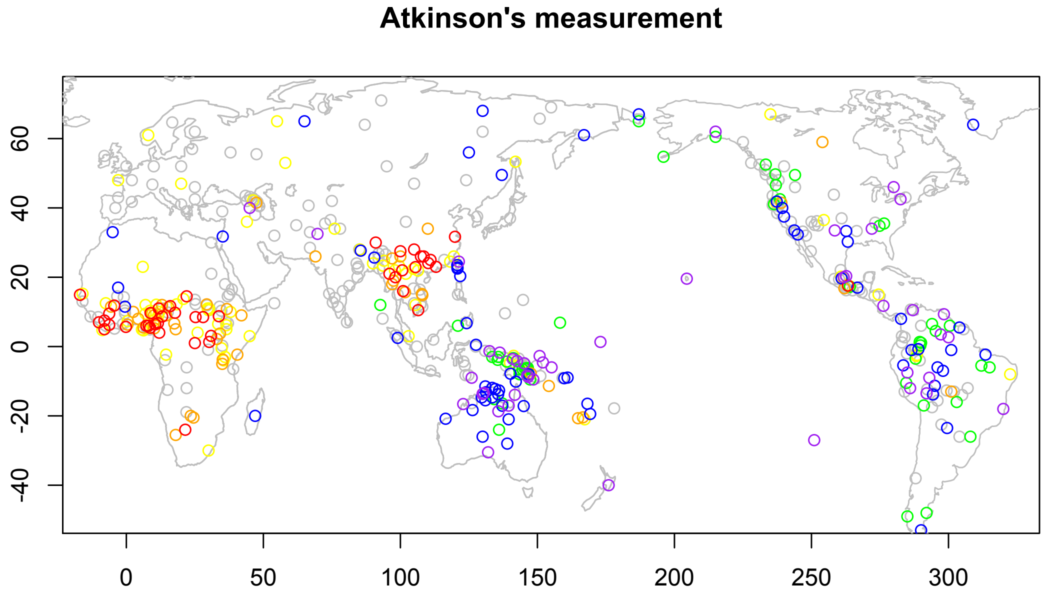


FIG. S2. Delaunay triangulation of 2519 languages from WALS, not crossing ancient water boundaries.

On a global scale, this distance measure produces very similar results to Atkinson's land distance (see Section 1.6), suggesting that they capture similar aspects of the linguistic reality in this context. However, our conception of language distance allows us to do local interpolations in a sensible manner. Just averaging over groups of languages in a circle of, say, a land distance of 100 kilometers will result in highly unequal groups depending on language density. In contrast, averaging over groups of languages within a distance of, say, maximally five "language crossings" results in much more balanced groups. With this distance it is possible to compute running averages for each sampled language *L* to show areal preferences. Basically, a maximum distance is chosen, and then the set of languages within this maximum distance is selected for each language *L*. An average is computed for all sampled languages within this set around *L* (note that the number of sampled languages is normally much smaller than the total 2519 languages in the Delaunay triangulation). This average is then plotted instead of the original value of *L*.

The first illustration in Fig. S3 shows the raw values of Atkinson's measure of phoneme inventory size. Although there are visually some areal preferences discernible, there is still a large amount of regional variation. The second illustration in Fig. S3 shows the same data, but now interpolated over areas with a maximum distance of five languages. Here there are clearly two areas with large phoneme inventories in Africa and Southeast Asia. Note that this areal distribution is highly similar to the areal distribution of tone marking alone (*S8*), once again indicating that tone marking is overvalued in Atkinson's measurement of phoneme inventory size. In contrast, Fig. S4 shows exactly the same illustrations, but now made on the basis of the UPSID data. The raw frequencies show even more variation, but the interpolation over areas of maximally a five-language distance clearly shows various areas with on average large phoneme inventories, viz. South Africa, the Caucasus, Northwest America, and minor clusters in Western Europe and Southeast Asia. These clusters exactly match linguistic intuitions about where languages with large phoneme inventories are to be found. Predominantly small phoneme inventories are found in New Guinea, Australia and South America, i.e. the furthest regions from Africa. No obvious origin discernible, as basically all of Africa, Europe, Asia and Northwest America show areas with large phoneme inventories.



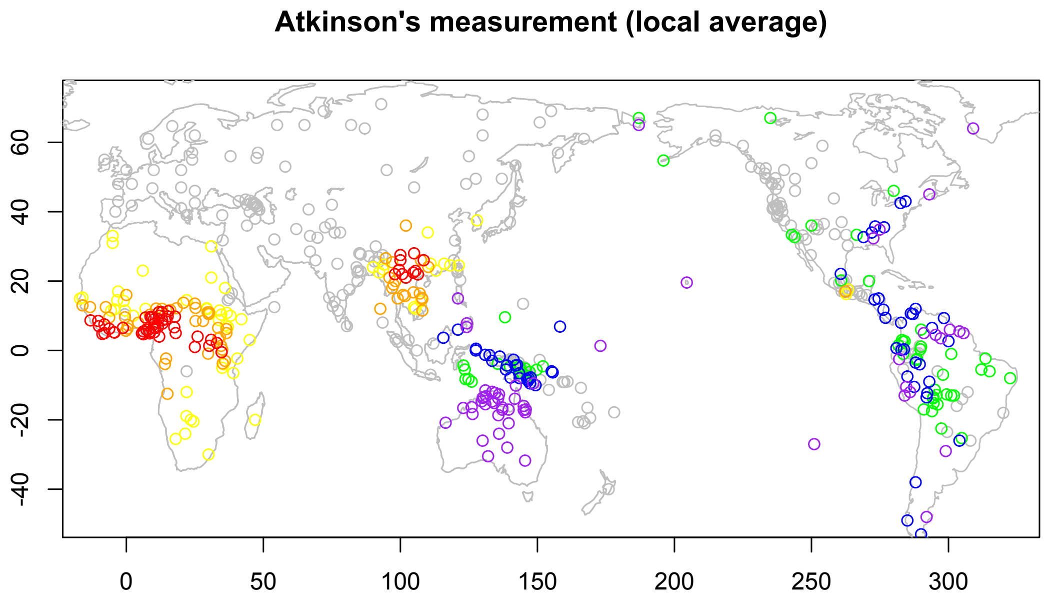
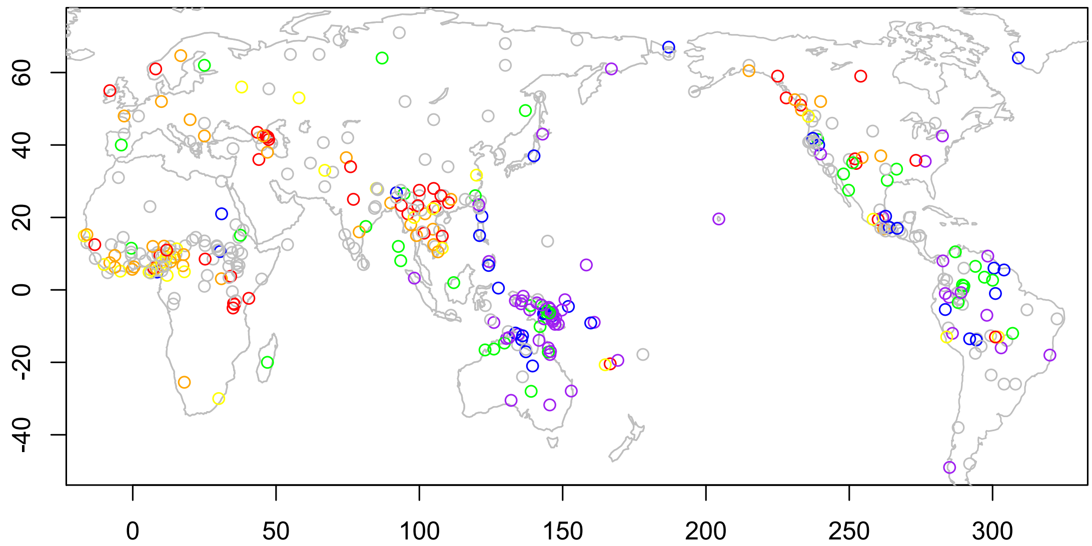


FIG. S3. Geographic distribution of phoneme inventory sizes according to Atkinson's measurement. Red/orange/yellow are the upper 10/20/30% of the sizes; purple/blue/green are the lower 10/20/30% of the sizes. The first plot shows the raw numbers, while the second plot shows for each language the local average, averaging over all languages sampled within a range of maximally a five-language distance. Clearly visible are two main regions with large phoneme inventories: Africa and Southeast Asia.

UPSID frequencies



UPSID local average

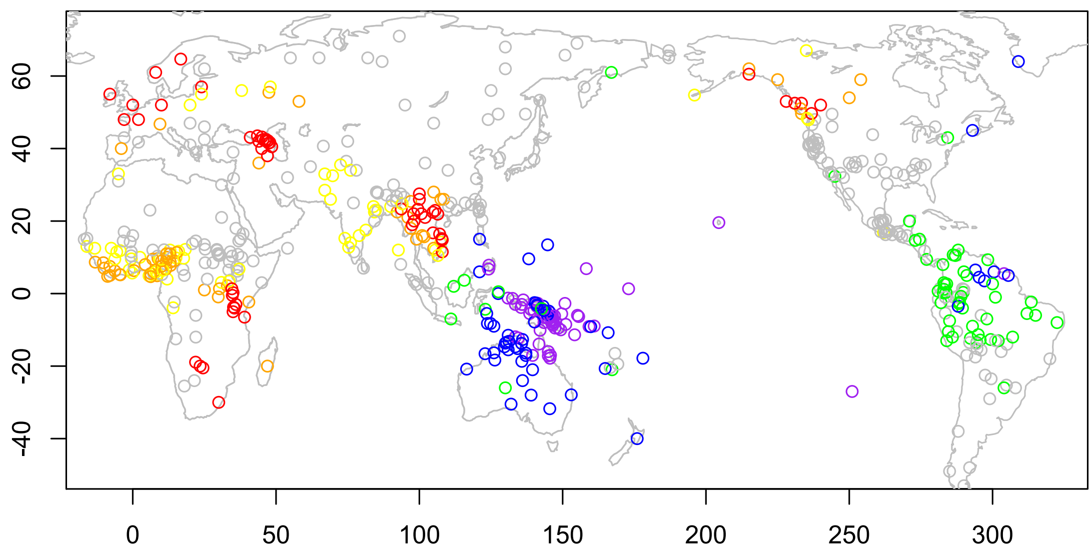


FIG. S4. Geographic distribution of phoneme inventory sizes according to the UPSID count. Red/orange/yellow are the upper 10/20/30% of the sizes; purple/blue/green are the lower 10/20/30% of the sizes. The first plot shows the raw numbers, while the second plot shows for each language the local average, averaging over all languages sampled within a range of maximally a five-language distance. There appears to be many more clusters of languages with an average high phoneme inventory as in Atkinson's measurement. Centers of high phoneme counts are attested in South and East Africa, the Caucasus, Western Europa, Southeast Asia and Northwest America.

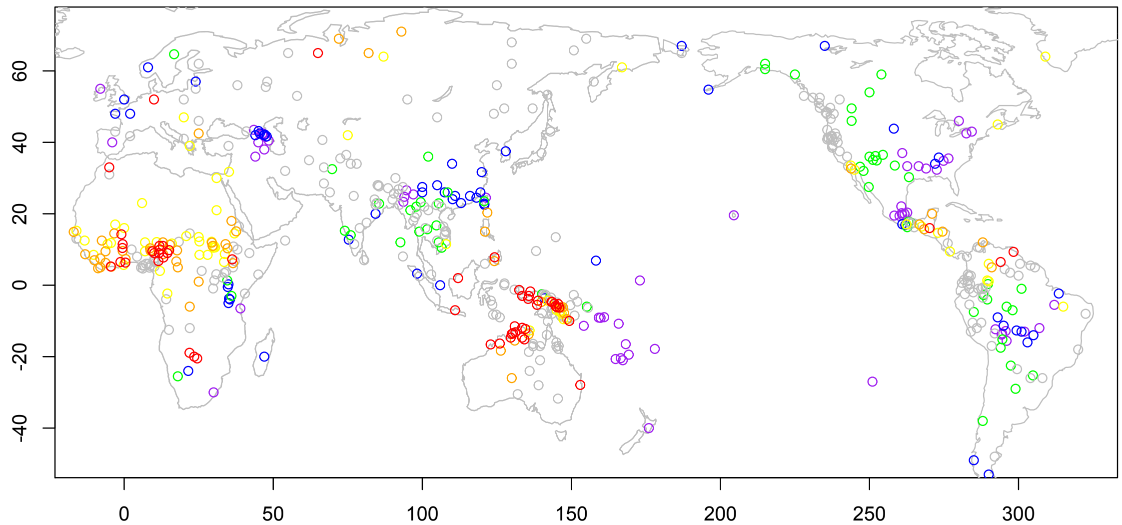


FIG. S5. Standard deviation of the UPSID phoneme inventory sizes (in log10) established for each language by taking all available sampled languages within a maximal distance of 5 languages. Red/orange/yellow are the lower 10/20/30% of the standard deviations (i.e. low variation); purple/blue/green are the upper 10/20/30% of the standard deviations (i.e. high variation). Africa and New Guinea/Australia show the least variation in their inventory sizes.

Instead of locally averaging over the actual number of phonemes, it is also highly informative to investigate the standard deviation of inventory sizes within local areas. Fig. S5 shows the standard deviation in inventory sizes from UPSID within a maximal distance of five languages at each language location. Africa and New Guinea/Australia are the predominant areas with little variation in inventory sizes. From the perspective of a serial founder effect, the low variation in New Guinea/Australia is exactly as would be expected, but the low variation in Africa, the supposed origin, is unexpected.

1.4. Correlation with speaker community size

It has repeatedly been observed that there is a positive correlation between the phoneme inventory size of a language and the speaker community size (*S17-S19*). Atkinson reiterates this observation and we can reproduce it also using UPSID ($r = 0.30$, $p = 7.18 \cdot 10^{-10}$, using data from *S16* for the population sizes). Note that for this correlation, we used the logarithm of population size and the logarithm of the phoneme inventory size. The analysis of the expected distribution of phoneme inventory size is still not settled (*S20*-*S22*), but using a logarithm seems to be preferable to using the raw numbers. However, this correlation shows strong dependency on the specific measurement of phoneme inventory size that is used. Scatter plots for various measurements are shown in Fig. S6 with a smooth spline indicating the local direction of the correlation. Globally, these correlations are all significant. However, the most important difference between these correlations is the behavior with small speaker communities. Atkinson argues that there is also a significant correlation "when the analysis is restricted to languages with speaker populations of 5000 or less, a range in line with speaker populations of modern hunter-gatherers" (Fig. S1 in *S1*). This significant correlation for small populations is crucial for Atkinson's proposal of a serial founder effect, because the founding populations would have been small. Unfortunately, the correlation for populations below 5000 is not significant at all with the other measurements of phoneme inventory (Weighted WALS: r = -0.04, p = 0.69, UPSID: r = 0.04, p = 0.64). With both these measurements, the correlation only reaches significance at the 5% level when much larger populations are included (all populations up to $5.0 \cdot 10^5$ for Weighted WALS, or $1.0 \cdot 10^5$ for UPSID), but such sizes are clearly outside the range of founding populations during the colonization of the world (*S23*).

FIG. S6. Scatter plots of speaker community size against phoneme inventory size for three different measurements of phoneme inventories with a smooth spline to show the local correlation effects. All correlations are significant over the whole population range, but when restricted to small speaker communities (up to 50,000 speakers) only Atkinson's measurement shows significance.

Nevertheless, the global correlation between speaker community size and phoneme inventory size is small but solid, though it is still far from clear how to explain it. We will here simply accept the correlation as given, and assume that it is not an accidental effect. Given the existence of this correlation, there is the question of the direction of causation. Whatever the reason for the correlation, it seems clear that it has to be the population size that has some kind of influence on language structure. It is highly unlikely that language structure influences population size, i.e. that languages with more phonemes favor the development of larger speaker populations. Further, the existence of large speaker populations (which we roughly define here as populations larger than 10^5 speakers) is probably a relatively recent phenomenon (*S23*), meaning that the correlation is most probably an effect that only arose after the human settlement of the world was already finished. Finally, the reason for a speaker population to grow large has various geographic, climatic, technological and sociopolitical reasons that are completely independent of the specific language being spoken, i.e. from a linguistic perspective it is pure chance that it happened to be language X that grew large instead of its neighbor Y (*S24*-*S25*)

Given this perspective, speaker community size is a factor to account for in the measurement of inventory size. The more so as the geographic distribution of large speaker communities is not random at all. There is a strong bias of large speaker communities to occur in Africa, Eurasia and Southeast Asia. Fig. S7 shows the geographic distribution of speaker community size, showing for each language the average if its own population size combined with the population sizes of the directly neighboring languages. The geographic bias is striking. Most importantly, assuming some kind of causal role of population size in determining phoneme inventory size, this geographic distribution of large speaker communities influences the geographic distribution of phoneme inventory sizes, favoring Africa, Europe and South Asia as being a region with large phoneme systems. So, the factor speaker community size has to be statistically removed when the distribution of phoneme systems across the world's languages is investigated.

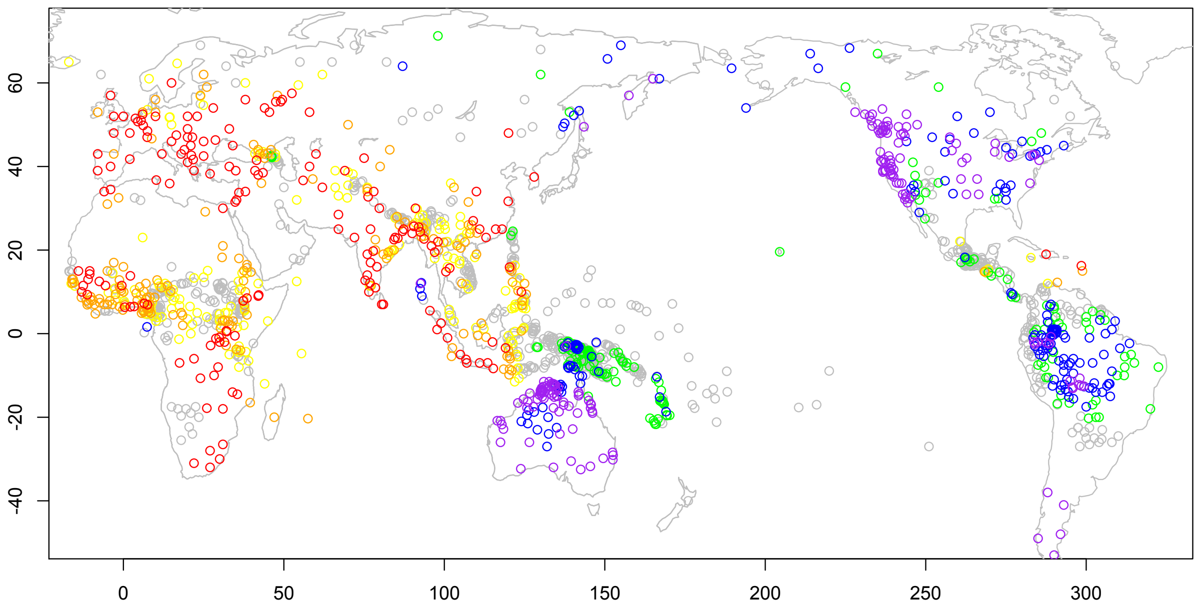


FIG. S7. Geographic distribution of speaker community sizes. For each language, the average population size for the language itself together with its direct neighbors is shown. Red/orange/yellow are the upper 30% of the population sizes; purple/blue/green are the lower 30% of the sizes. Extremely small (and often highly endangered) languages predominate in Australia and North America, while there are many small languages in New Guinea and South America. Languages with large speaker communities predominate in Europe, Africa, South Asia, and Southeast Asia.

1.5. Distribution over macroareas

As an approximate indication that there might be an 'out-of-Africa' effect in the geographic distribution of phoneme inventory sizes, Atkinson presents a boxplot in the original article comparing the inventory sizes across six macroareas as distinguished in WALS (Fig. 1B in S*1*). This boxplot is replicated here in Fig. S8, top left, where it is compared to other measurements of phoneme inventory size. However, there are various problems with this boxplot.

Atkinson does not elucidate the definition of the macroareas, but from the visual inspection of his boxplot it very much looks like he took the definitions of macroareas as available in WALS. It is important to realize that the boundaries of the macroareas in WALS have rather special definitions (*S26*). They were defined to be linguistically maximally independent from each other and they were never intended to be used to investigate the peopling of the globe. The geographic distribution of these macroareas is shown in Fig. S9. Specifically, the relative order of Eurasia and Southeast Asia is difficult to interpret from a viewpoint of ancient human population movements. For reasons of comparability, we have retained Atkinson's order of macroareas in all of our boxplots in Fig. S8: Africa - Southeast Asia - Eurasia - North America - South America - Oceania.

Further, the term 'Oceania' as used by Atkinson in his boxplot does not seem to be appropriate. The area 'Oceania' does not exist in WALS, but there is an area 'New Guinea and Australia' that matches the numerical distribution in the boxplot. Linguistically, this difference is crucial, because Oceania would basically represent a grouping of languages from New Guinea and Australia together with the Austronesian family of languages. The Austronesian languages only dispersed relatively recently into the Pacific region (starting about 4,000 years ago, *S27*), while the non-Austronesian languages from New Guinea and Australia already populated this area long before the Austronesians (possibly even dating back to the original peopling of the globe). Thus, we decided to change the label in our boxplots to the more appropriate "New Guinea and Australia" (abbreviated 'NG+Aus').

There are six different versions of the boxplot shown in Fig. S8. The boxplots differ depending on which data is used and whether to account for population size or not. In all plots, South America and New Guinea/Australia seem to be substantially lower than the other areas, with Southeast Asia being mostly intermediate. Africa, Eurasia and North America are approximately equally high in all plots. A preference for Africa is only found in one of the boxplots, viz. the one replicating Atkinson's method. In contrast, North America shows the highest averages when using the data from UPSID and regressing to population size. A clear 'out-of-Africa' cline is thus only discernible using the exact details of Atkinson's approach. Slight variations in the method of measurement do not suggest this effect.

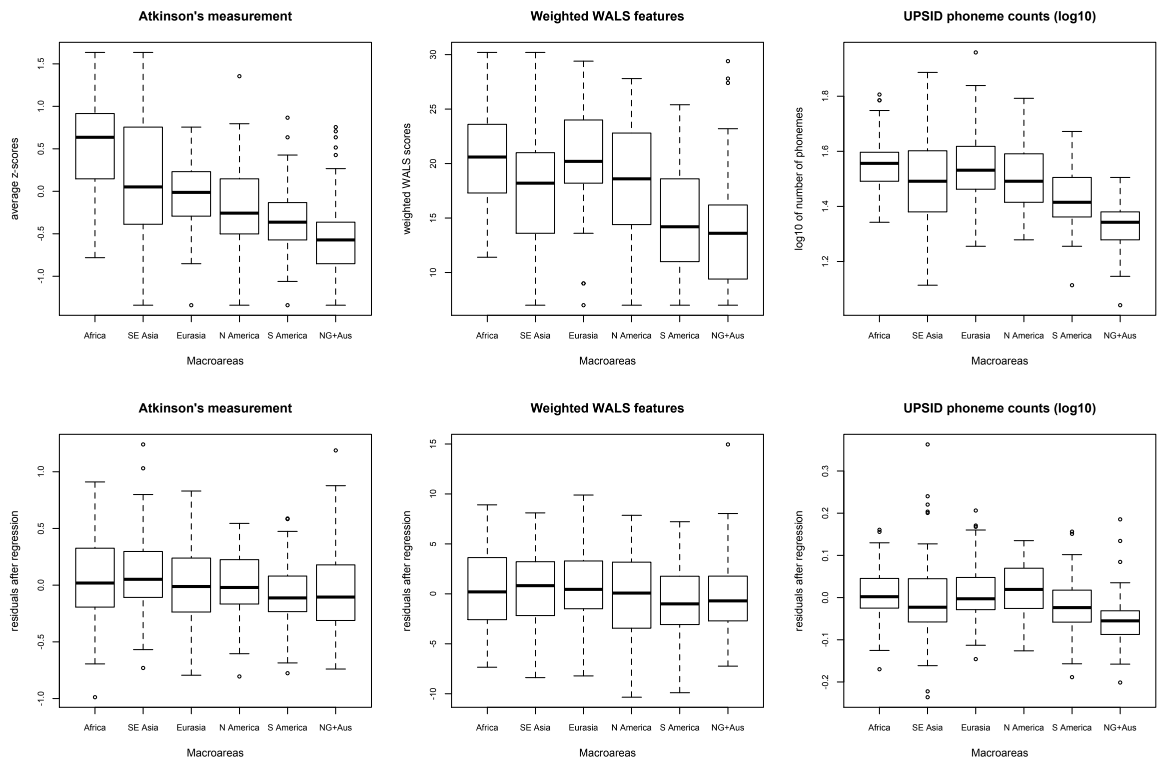


FIG. S8. Boxplots showing phoneme inventory size by macroarea. The upper row reports the raw numbers of phoneme inventory size, while the lower set of boxplots reports the residuals after regressing to population size, including linguistic genera as a random factor. The leftmost boxplots use Atkinson's measure of phoneme size (average z-scores of three WALS features). The middle boxplots use the weighted sum of the same WALS features, and the rightmost boxplots use the phoneme counts from UPSID. All variants show a relatively small phoneme inventory for South America and New Guinea/Australia. The exceptionally large phoneme inventories for Africa are only attested in Atkinson's original measures (this boxplot was printed in his original article). The residuals from UPSID (our favored measure) show North America as the macroarea with the highest phoneme inventories.

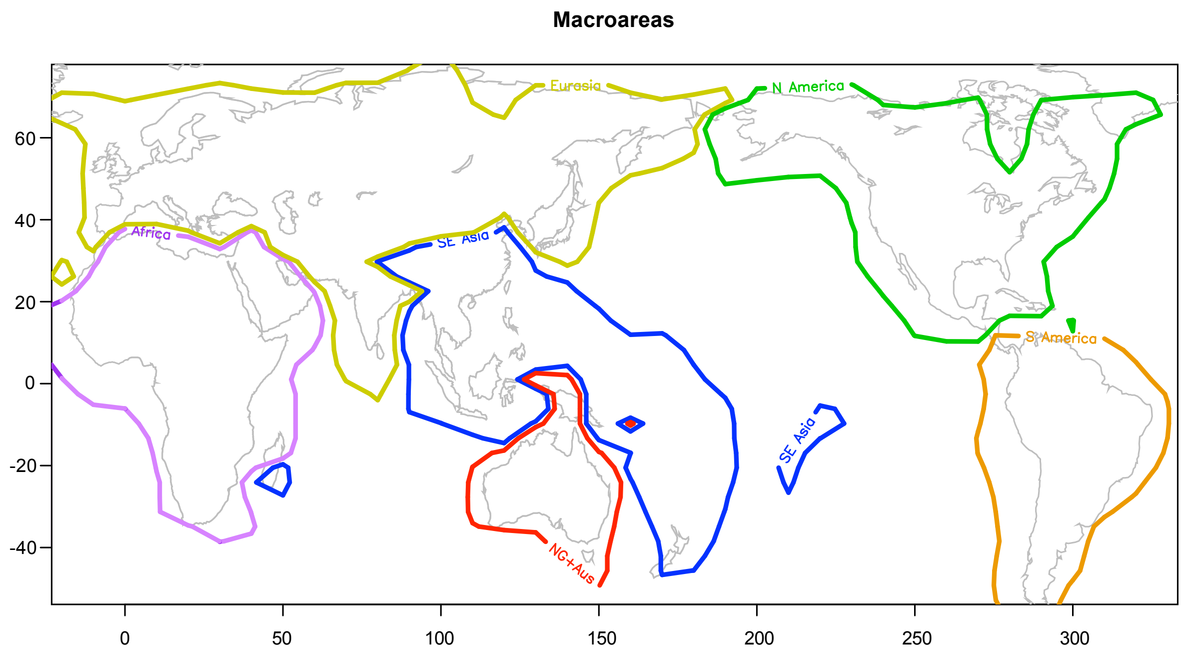


FIG. S9. Geographic distribution of WALS macroareas.

1.6. Global clines of phoneme inventory size

Atkinson's research clearly was inspired by previous work investigating human evolution, which found clines of decreasing genetic and phenotypic diversity in modern humans the farther away from Africa the sampled populations are (*S28-S29*). Basically, in this method the trait of interest (here, phoneme inventory size) is measured at several geographic locations and then regressed on the distance to a given geographic origin, while controlling for various possible confounds such as population size. Then, several possible such origins are considered and the Bayesian Information Criterion (BIC; *S30*) of these regression models is computed. These possible origins are then sorted in order of increasing BIC. The origin with the minimum BIC is considered to give the best relationship of the trait of interest to geographic distance and taken to be the most probable ("true") origin of expansion. Please note that at this stage neither the sign nor the size of the regression coefficient of geographic distance are considered. This best fitting model could be one with decreasing or increasing trait values as a function of distance. Next, Atkinson selects those locations at most 4 BIC units away from this optimum as having 'considerable support' in being the origin of the expansion. Please see section 1.8 for a detailed critique and analysis of this method.

We replicated this method used by Atkinson to assess the global origin of phoneme inventory size as follows. Following Atkinson, we considered in turn each of the available languages as a possible origin, and we regressed the phoneme inventory size of all other languages on the geographic distance to the considered origin, while controlling for speaker community size as a second regressor and dealing with the genealogical non-independence of languages by including linguistic genera as a random factor. An alternative model not considered by Atkinson is a quadratic factor in geographic distance, which turned out to change the results drastically (see below). Further, we took the languages within 4 BIC points from the optimum model as the probable region of origin. Note that in all these regressions, we always took the logarithm of the speaker community size. Likewise, we used logarithms of the UPSID counts (cf. Section 1.4), but not for the other measurements of phoneme inventory size.

The geographic distribution of the languages within the BIC+2,4,6,8 range is shown to the left in Fig. S10. Shown in blue is the BIC cluster according to Atkinson's rough measurement of phoneme inventory size. This cluster shows exactly the West African origin as claimed by Atkinson. Shown in green is the BIC cluster on the basis of the weighted sum of the same WALS features. This green cluster is still based in Africa, but shows a markedly different geographic orientation, being centered on *Sandawe* in eastern Africa. Shown in red is the BIC range based on the UPSID counts of phonemes per language. This BIC range actually consists of two clusters. The minimum BIC is attested for the East African languages *Sandawe*, but the second lowest BIC is found for the Caucasian language *Hunzib*. Clusters of low BIC values arise around these two centers, which only merge into a single cluster when BIC-values above BIC+3 are included. When we added a quadratic geographic term into the regression model (shown to right in Fig. S10), there was no change to the BIC area according to Atkinson's measure. However, for the other two measurements of phoneme inventory size the clusters of minimal BIC languages shifted dramatically to the eastern tip of New Guinea. In this model, the origin of phoneme inventory size consists of languages with small phoneme inventories.

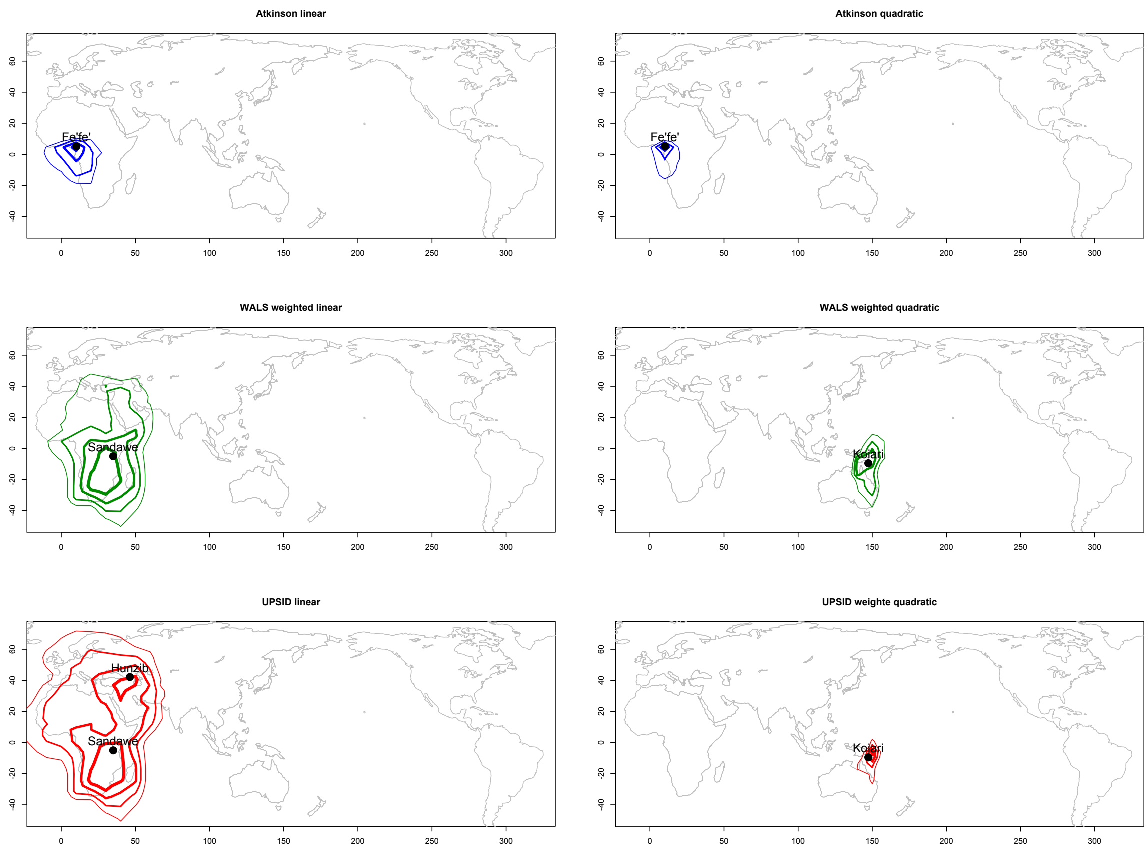


FIG. S10. Geographic distribution of languages within the BIC+2,4,6,8 range, indicated with contours of diminishing thickness. In blue is Atkinson's own measurement of phoneme inventory size, showing his claimed West African origin. In green is the weighted WALS assessment of phoneme inventory size with an East African origin. In red is the phoneme inventory count in UPSID with a double East African and Caucasian origin. To the right the models with an additional quadratic geographic term are shown. The blue area does not change in the quadratic model, but the two other measurements now show New Guinean origins.

In summary, using the UPSID data results in a second origin of large phoneme inventories outside Africa (in the Caucasus), and in general the size of the BIC+4 cluster is markedly larger than the BIC+4 cluster based on Atkinson' data. However, the 'true origin' of phoneme inventory size is still in Africa, while the Caucasus could possibly be construed as a very ancient secondary development. In contrast, when we add a quadratic geographic term to the regression, the supposed 'origin' is placed in New Guinea, and the original state of the phoneme inventory size would be one with small inventories.

1.7. Global clines of other WALS features

The explanation presented by Atkinson for the African origin of large phoneme inventories (i.e. a serial founder effect in which small daughter populations lost linguistic categories) is general enough that it should also hold for other linguistic characteristics that involve some kind of 'more' vs. 'less' explicit marking structure. Note that we will refer to this 'more' vs. 'less' explicit marking as 'complexity' here, even though the definition of complexity in language is a hotly debated topic (*S31-S33*).

Contrary to the general explanatory principles proposed by Atkinson, other applicable WALS features do not indicate the same scenario as implied by Atkinson's explanation. We investigated an ad-hoc selection of 16 WALS features that are easily construed as involving some kind of structural complexity difference. All these features distinguish between languages that have some kind of overt morpho-phonological marking vs. languages that do not have any overt linguistic marking structure (which normally means that this second group of languages use other, more implicit, strategies to express the same content). For all these characteristics we replicated the same analysis as used by Atkinson. The first problem for Atkinson's explanation is that we find 'origins' all over the world, not just in Africa. And second, the implied original linguistic state can go both ways, being either the one with the most or with the least explicitly marked structures. These 16 features are the following (see also Fig. S11):

- WALS 9 “The velar nasal” (*S34*) describes the usage of the velar nasal consonant. Most languages do not have such a phonemic consonant, some have such a phoneme, but it can only be used in restricted non-initial environments. Finally, a large set of languages allows the velar nasal also in initial position. The minimal BIC is found in *Madurese* Southeast Asia & Oceania) with only 6 languages (1% of all sampled languages) being within the BIC+4 range. This area typically has unrestricted usage of the velar nasal (i.e. more structure).
- WALS 10 “Vowel nasalisation” (*S15*) describes whether a language has phonemic vowel nasalization or not. The minimal BIC is found in *Maybrat* (Australia - New Guinea) with 46 languages (19%) within the BIC+4 range. These languages typically do not have nasalization (i.e. less structure).
- WALS 12 “Syllable structure” (*S35*) classifies the complexity of syllable structures. The minimal BIC is found in *Yupik* (on the Eurasian - North American border), but a secondary center is English, showing two disconnected areas within the BIC+4 range with in total 13 languages (3%). The whole of Eurasia and large parts of North America are characterized by complex syllable structures (i.e. more structure).
- WALS 22 “Inflectional synthesis of the verb” (*S36*) describes how many inflectional categories are marked on a verb in the languages investigated. The minimal BIC is found in *Koasati* (North America), an area that typically has high inflectional synthesis. However, all 145 samples languages fall within the BIC+4 range, so this characteristic does not show any clear founder structure.
- WALS 27 “Reduplication” (*S37*) describes the extent to which languages use reduplication. The minimal BIC is located in *Uradhi* (Australia) with a large area of 103 languages (28%) around it within the BIC+4 range. These languages typically have productive full and partial reduplication (i.e. more structure).
- WALS 30 “Number of genders” (*S38*) describes how many grammatical genders are distinguished in languages. The minimal BIC is attested in *Cocopa* (North America) with 33 neighboring languages (13%) within the BIC+4 range. These languages typically do not have any grammatical gender (i.e. less structure).

- WALS 34 “Occurrence of nominal plurality” (*S39*) describes the extent to which languages use overt plural marking on nouns. The minimal BIC is found in *Greek* (Eurasia) with 48 languages (16%) in Europe and northern Africa within the BIC+4 range. These languages typically have obligatory plural marking on all nouns (i.e. more structure).
- WALS 41 “Distance contrasts in demonstratives” (*S40*) describes how many distance contrasts languages mark in their demonstratives. The minimal BIC is located at *German* (Europe) with a large area of 94 languages (40%) being within the BIC+4 range. These languages span an enormous area, basically including all of Eurasia, Mainland South and Southeast Asia, and large parts of Africa. The languages in this area typically have just a two-way demonstrative system (i.e. less structure).
- WALS 47 “Intensifiers and reflexive pronouns” (*S41*) describes whether languages have a specific intensifier, or whether they simply use the reflexive pronouns for this function. The minimal BIC is found in *Alamblak* (New Guinea) with an area of 30 surrounding languages (18%) within the BIC+4 range. These languages typically do not have specialized intensifiers (i.e. less structure).
- WALS 49 “Number of cases” (*S42*) shows how many noun cases a language distinguishes. The minimal BIC is located in *Russian* (Eurasia), but the set of languages within the BIC+4 range includes all 261 languages sampled, so this characteristic does not show any clear founder structure.
- WALS 55 “Numerical classifiers” (*S43*) describes whether languages use numerical classifiers. The minimal BIC is attested in the Southeast Asian language *Loven* with a small area of 11 languages (3%) being within the BIC+4 range. These languages typically have obligatorily usage of numeral classifiers (i.e. more structure).
- WALS 59 “Possessive classification” (*S44*) describes how many noun classes are distinguished in the formal marking of pronominal possession. The minimal BIC is attested in *Chuchki* (Eurasia, on the boundary to North America) with an extremely large group of 72 languages (30%) within the BIC+4 range, spanning over all of Eurasia and North America.

These languages typically do not have any noun class distinctions for pronominal possession (i.e. less structure).

- WALS 65 "Perfective/imperfective marking" (*S45*) classifies languages according to whether they grammatically mark a perfective/imperfective distinction or not. The minimal BIC is found in *Kanakuru* (Africa) with a large group of 68 languages (31%) within the BIC+4 range, spanning all of Africa, the Near East, and parts of Europe. These languages typically grammatically make such a distinction (i.e. more structure).
- WALS 67 "The future tense" (*S46*) shows which languages have specialized grammatical marking to indicate future tense. The minimal BIC is located in *Palaung* (Oceania), but the set of languages within the BIC+4 range includes all 222 languages sampled, so this characteristic does not show any clear founder structure.
- WALS 77 "Semantic distinctions of evidentiality" (*S47*) describes whether languages grammatically mark evidentiality or not. The minimal BIC is found in *Kashaya* (North America) with a large group of 119 languages (28%) within the BIC+4 range, spanning all of North America. These languages typically have grammatical evidentials (i.e. more structure).
- WALS 107 "Passive constructions" (*S48*) described whether languages have a specialized passive construction or not. The minimal BIC is found in *Karok* (North America) with a group of 60 languages (16%) being with in the BIC+4 range. Within this range, a secondary center is identified around *Uradhi* (Australia). The North American group typically has a specialized passive, while the Australian groups typically does not.

Thus, for three features there is no clear origin at all, five have either a very large or more than one 'origin' area(s), and for the remainder we found origins all around the globe (except for South America). Moreover, seven features have an 'origin' showing more structure (i.e., the cline is one of decreasing trait value), while five show the opposite pattern, with an 'origin' showing less structure (i.e. a cline of increasing trait value). These results show that, from a purely statistical point of view, the African origin of high phonemic inventory size is just one of many possible origins of linguistic diversity as identified by Atkinson's approach. The only other similar origin is attested for WALS 65, for which the analysis suggests an African origin for explicit perfective/imperfective marking. We do not see any reason why phoneme inventories or perfective/imperfective marking would be more telling for linguistic origins than any of the other parameters discussed here.

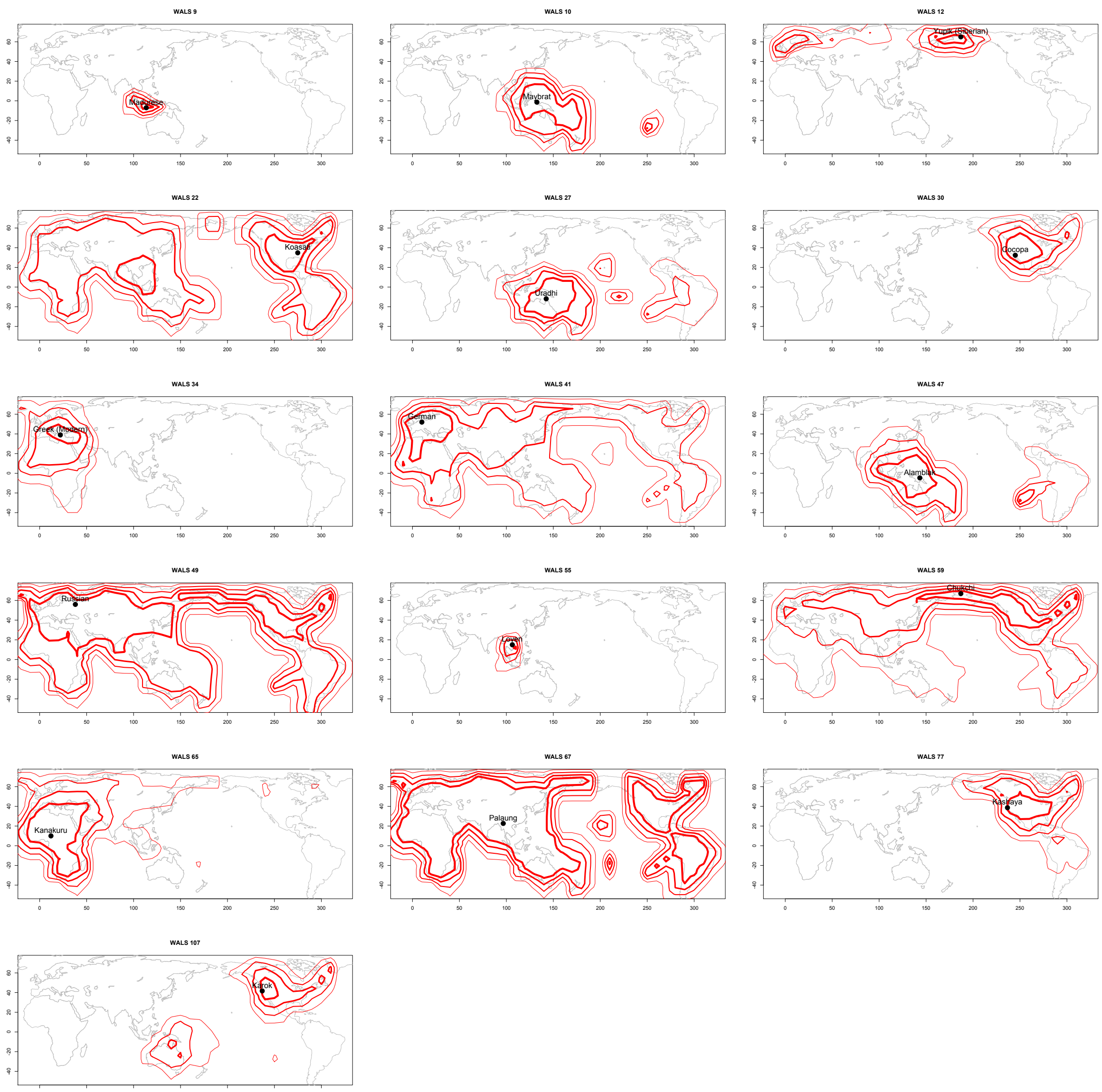


FIG. S11. Searching for the 'origin' of 16 WALS features. The BIC-minimum is indicated with a black dot including the name of the language. The geographic range of the languages within the BIC+2,4,6,8 range is indicated with a red contour with diminishing thickness. The origins occur all over the globe, except for South America.

1.8. Searching for an origin: analysis of Atkinson's BIC-based methodology

The methodological crux of Atkinson's paper is the search for the 'origin' that optimizes the regression of phonological inventory size on the geographic distance to that origin. This is practically implemented as an optimization procedure over a finite set of such 'origins' (the locations of all languages in the sample) where the objective function over the origins is the Bayesian Information Criterion (BIC; *S30*) of the regression model. Thus, if we denote a possible origin as *x*, the set of such possible origins as *X*, and the objective function as *f(x)*, we have:

(3) $$f(x) = BIC(p \sim d(x) + s + (1|g))$$

where we used the "R notation" for mixed-effects models (*S49*), namely regressing the phoneme inventory size *p* of the languages on the distance from those languages to the origin *d(x)* while controlling on the languages' speaker population sizes *s* and taking into account the non-independence between related languages by taking the genus *g* as a random effect. With these, Atkinson's method first searches for the "best origin" x_0 such that $f(x_0) = min\, f(x)$. Next, he selects the subset of origins $O = \{x_1, x_2, \dots x_n\} \subseteq X$ such that $f(x_i) \leq f(x_0) + t$, where *t* is an a priori fixed threshold. With these notations, the general formula of BIC is

(4) $$f(x) = -2 \cdot ln(L(x)) + k(x) \cdot ln(n(x))$$

where *ln(.)* is the natural logarithm, *L(x)* is the regression model's maximum likelihood, *k(x)* is the model's number of free parameters and *n(x)* is the number of observations for the model. Now, in this search procedure both *k(x)* and *n(x)* have fixed values *k* and *n*, respectively, so that we have the reduced equation:

(5) $$f(x) = -2 \cdot ln(L(x)) + k \cdot ln(n)$$

where only the maximum likelihood varies between origins. Thus, the search for the set of "best origins" *O* reduces to finding those *x* such that the difference:

(6) $$f(x) - f(x_0) = -2 \cdot ln\big(L(x)\big) + 2 \cdot ln\big(L(x_0)\big) = 2 \cdot \ln\left(\frac{L(x_0)}{L(x)}\right)$$

is smaller than *t*, which reduces further to:

(7) $$\frac{L(x_0)}{L(x)} \leq e^{t\backslash 2}\ .$$

Atkinson's threshold *t* is 4 units, which means that he selects those "origins" *x* which have a maximum likelihood *L(x)* at most $e^{4/2} = e^2 \approx 7.4$ times smaller than that of the best fitting origin, which, in plain English, means that he would select those "origins" at most 7.4 times less likelier than the best one.

It remains somewhat unclear from the Atkinson's paper where this threshold of 4 BIC units comes from. Atkinson (*S1*) cites a paper by Andrea Manica and colleagues (*S28*) which applied the same methodology to phenotypic variation and which, in turn, cites (p. 349) a book by Burnham and Anderson (*S50*) but no page is given. Fortunately, the same group has published another paper using the same methodology (*S29*) where, as a justification for the same threshold, they cite instead (p. 810 in *S51*) a paper dealing with model selection in capture–recapture studies using Akaike's Information Criterion, AIC, but we were unable to find any mention in this paper of BIC or the 4 units threshold for providing "considerable support" (*S28*-*S29*). However, we did manage to find on page 70 in (*S50*) "some rough rules of thumb" concerning model selection using again only AIC (and variants), but which "are particularly useful for nested models". Besides the fact that the models tested by Manica and colleagues (and, by extension, Atkinson) are not nested, the "rough rules of thumb" are: a difference of 0-2 units gives "substantial level of empirical support", 4-7 give "considerably less", while >10 gives "essentially none". Differences in AIC (which is defined as

$-2 \cdot ln(L(x)) + 2 \cdot k$) reduced very similarly to BIC (as we have demonstrated above) to maximum likelihood ratios due to constant numbers of free parameters and observations, probably allowing the application of AIC rules of thumbs also with BIC values. However, we are still facing the issue of non-nestedness and the fact that 2 (and not 4) is the recommended threshold for “substantial support”. Using 2 instead of 4 of course substantially reduces the size of the regions of “origin” due to the reduction in the number of selected locations in *O*.

Another property of this procedure of using a threshold *t* in selecting the set of origins *O* is that it necessarily results in contiguous geographic regions around the best fitting location x_0. This is very dramatically illustrated by using various types of randomization: (a) we can randomly shuffle the actual phoneme inventory sizes around the languages or (b) we can generate random numbers from a standard normal distribution as the values of the phoneme inventory sizes or (c) we can add random normal noise of given standard deviation to the distances between languages (Fig. S12). In all these randomizations the method still finds a geographically contiguous ‘origin’. This shows that the method necessarily finds contiguous spatial regions, which might be seen as a strength if one thinks that the assumptions are justified, but can also be taken as a caveat against seeing too much in the geographic continuity of the “regions of origin”. Thus, for randomly permuted data (a) and purely random data (b) the BIC optimization method and a threshold $t = 2$ produces “strongly supported” regions of origin even if there is no such thing in the processes generating the data! The only distinction between these random(ized) data sets and the “real” set is that the $t = 4$ usually includes the whole world. Thus, it seems that it is somehow important how strongly restricted the $t = 4$ cluster is in making claims about the “origin” region.

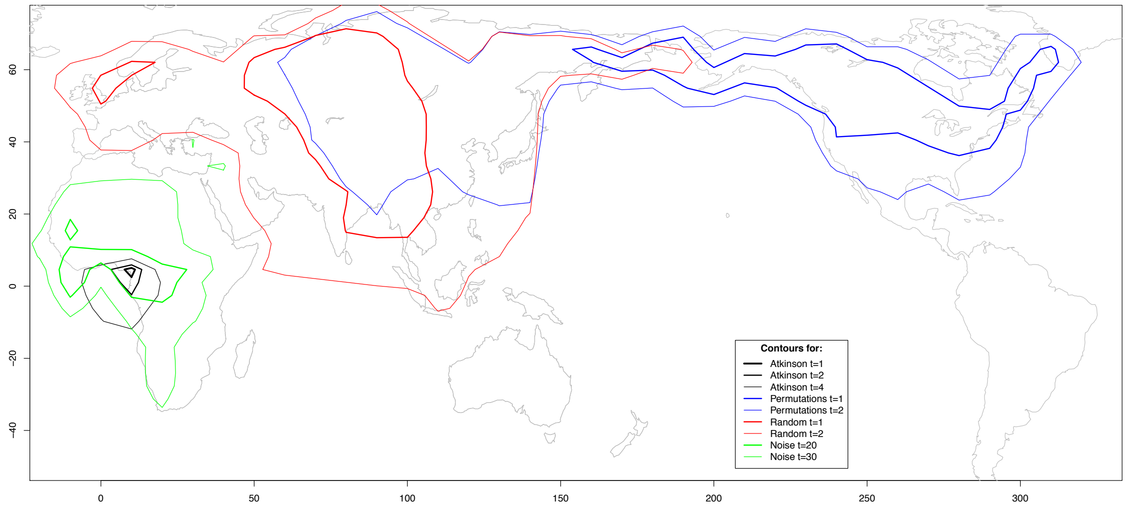


FIG. S12. Illustrative regions identified as "origin" by the BIC+t units method for different types of random(ized) data and thresholds, t. In black, the original Atkinson origins for t=1, 2 and 4. In blue, the phoneme inventory sizes have been randomly permuted across languages destroying any geographic information (case a) for $t = 1 \; or \; 2$ ($t = 4$ encloses the whole world). In red, the language phoneme inventory sizes are in fact random numbers from a normal distribution with mean and standard deviation matching the Atkinson data (case b) for $t = 1 \; or \; 2$ ($t = 4$ encloses the whole world). In green, we added random noise from a normal distribution with mean 0 and standard distribution 4 to the geographic distances (case c) for $t = 20 \; or \; 30$ (smaller t results in a very small region in West Africa).

Two final points of criticism of this method are that, first, it explicitly searches for a linear model and, second, assumes a single geographic origin. The first issue is illustrated by the UPSID data where adding a quadratic distance factor to the model gives a better fit ($min \; BIC_{linear} = 178.89$ compared to $\min BIC_{quadratic} = 134.36$) and suggests an origin in Australia/New Guinea (Fig. S10). For the second issue we can only sketch a verbal criticism here leaving the actual quantification for future work. The issue is more general and concerns many claims in research conducting apparently crucial hypothesis testing experiments, which test two competing hypotheses and end by unambiguously rejecting one in favor of the other. Unfortunately, sometimes either (a) one of the hypotheses completely fails to be included in the universe of hypotheses considered or (b) a highly distorted version of it (what could be seen as a “straw man” hypothesis) is proposed instead.

The paper by Atkinson suffers from both shortcomings because (a) the method that is used to search for origins by minimizing BIC cannot deal with other, more appropriate scenarios including massive horizontal transfer through language contact and multiple waves of migration. Thus, the BIC minimization really only selects the best hypothesis from within a homogeneous set of similar hypotheses, which all assume an underlyingly single expansion. Second (b), when Atkinson actually tests an alternative hypothesis (methodologically following *S28*) by adding a secondary origin into the model besides the best fitting origin in Africa (a model that is rejected), we are told that this eliminates the possibility of “language polygenesis” (*S1*, p. 347). We conjecture that nobody really entertains a model of “language polygenesis” in such a simplistic sense that it would leave traces detectable by Atkinson’s method.

1.9. Software packages used

All calculations in this paper were performed in *R* (*S52*), crucially using the following packages and functions:

- lme4 (*S49*): function *lmer* for mixed-effects models
- languageR (*S53*): function *pvals.fnc* for significances of mixed-effects models
- akima (*S54*): function *interp* for spatial interpolation
- deldir (*S55*): function *deldir* for Delaunay triangulation
- sna (*S56*): functions *geo.dist* for distances on a graph and *gplot* for graph plotting
- fields (*S57*): function *world* for plotting world outlines

2. Supporting Text

2.1. About the term ‘phonemic diversity’

Atkinson uses the term ‘phonemic diversity’ interchangeably with ‘phoneme inventory size’. In his usage, high phonemic diversity is equivalent to large phoneme inventory size, and vice versa. We consider this to be a rather unfortunate usage of the term ‘diversity’ in this context. Having more phonemes usually implies that the phonemes will be more similar among each other, as they make finer-grained distinctions within the same phonetic space. Thus, it would linguistically be more sensible to define phonemic diversity on the basis of the internal structure of the phonemic inventory and not simply on the basis of the number of distinctions.

Further, we believe the usage of the term ‘diversity’ in the current context indicates an underlying confusion. Atkinson’s approach was clearly inspired by previous work in biology on the genetic and phenotypic diversity in modern humans (*S28-S29*). However, in this biological work the concept ‘diversity’ refers to the quantification of the amount of variation within populations of individuals. In contrast, Atkinson’s linguistic diversity refers to differences between individual languages, not populations of languages. The proposed underlying parallelism seems to be that a structural property of language, such as the number of phonemes, is equivalent to a genetic locus in a population. However, this would imply that the, say, 23 different consonants in language *X* are somehow equivalent to the 23 different alleles in population *Y* or the 23 varieties a skull structure that can be found in the same population.

This strikes us as being a wrong parallel because of the different dynamic properties involved. For example, a basic process in population genetics is drift, whereby, in the absence of sources of genetic novelty, a population tends towards homogeneity through random sampling of genes across generations, leading to less diversity. However, there is no sense in which a language spontaneously becomes less diverse in this sense, i.e. necessarily reducing the number of phonemes through drift until becoming homogeneous in the limit, with ultimately a single segment left in its inventory. In contrast, population of language would become less diverse through this mechanism. However, when West Africa is interpreted as a population of languages, then this area is specifically low on variation, as we have shown in Fig. S5.

2.2. Stability of phoneme inventory size

In order to use phoneme inventory size with the goal to recover a signal dating back to the proposed out-of-Africa migration of modern humans 50-70,000 years ago, this aspect of human language must be stable enough to conserve this signal over this long stretch of time. 50-70,000 thousand years represent an enormous time span from a linguistic point of view, given that the comparative method in historical linguistics currently seems unable to go beyond approximately 10,000 years ago (*S58*). There are promising signs that newer methods based on typological features might go beyond that (*S59*) but it is currently unclear how much far back in time they can see.

We believe that the phonemic inventory size as construed by Atkinson does not have the required stability to preserve such a deep signal. One of the current authors has recently proposed an approach to measure typological stability based on Bayesian phylogenetic methods (*S60*, see also earlier work in *S61***)**. These measurements of stability have only relative values due to the absence of reliable calibration points for most language families. They simply represent relative stabilities among a large set of typological features from WALS across many families. He found that ‘tone’ (WALS 13) is one of the most phylogenetically stable features, ‘vowel quality inventory’ (WALS 2) is of average stability, while ‘consonant inventories’ (WALS 1) is one of the most unstable features. Thus, Atkinson’s composite of these features to represent the phoneme inventory size does not seem to be able to retain enough old information. Interestingly, other more stable features such as ‘vowel nasalisation’ (WALS 10) and ‘reduplication’ (WALS 27) produce non-African ‘origins’ (section 1.7 and Fig. S11).

2.3. About the serial founder effect in human evolution and language

A serial founder effect represents probably the simplest explanation for an observed global cline of decreasing diversity in genetic and phenotypic data (*S28,S29*). However, it is not the only possible explanation for such a cline. Alternative processes resulting in similar patterns are represented by isolation-by-distance genetic exchange dominated by Africa due to its long-term larger population size (*S62*), successive selective sweeps (*S63*) or multiple dispersals not necessarily all originating in Africa (*S64*). These proposals suggest more complex scenarios involving not only population movement and expansion (as the successive founder effect does) but also genetic exchange, admixture and possibly natural selection as well.

In the case of language, if a cline of decreasing phonemic diversity originating in Africa were true (which we doubt, as argued in Section 1.5 and 1.6), its interpretation solely in terms of a linguistic serial founder effect would be at least as simplistic as in genetics, even more so given the pervasive occurrence of horizontal transmission processes that are active in language (i.e. borrowing, mixing, super- and substrate effect; *S65,S66*). There are also other factors that correlate strongly with phoneme inventory size, like latitudinal distance from the equator (*S67*), suggesting that a single-origin model assumed by Atkinson is not necessary the only possible model to explain a world-wide cline. Further, the mechanism to account for a serial founder effect, as suggested by Atkinson, (namely that small daughter languages would lose phonemes in the process of splitting off) does not hold (see Section 1.4) nor is there any trace of a plausible mechanism for this known to us from the linguistic literature (*S68*). Also, other inventory-like linguistic phenomena did not follow the same process (see Section 1.7). All of this is in stark opposition to genetics, where mechanisms underlying a serial founder effect are well understood and widely attested.

In summary, we believe that the genetic findings are not a solid basis for Atkinson's metaphor of a linguistic serial founder effect. Thus, even if the decline from Africa in phonological inventory size were true, this would not make a serial founder effect the most obvious explanation.

3. Tables

The following data was compiled using basically data from WALS (*S4*) with the addition of UPSID counts (*S3, S10*). The population data are from (*S16*). All information is aligned using WALS codes.

CODE	*NAME*	*GENUS*	*LONG*	*LAT*	*AREA*	*POPULATION*	*W 1*	*W 2*	*W 13*	*UPSID*
abi	Abipón	Guaicuruan	-61	-29	South America	NA	5	2	NA	20
abk	Abkhaz	Northwest Caucasian	41	43.08	Eurasia	105952	1	1	1	NA
ach	Aché	Tupi-Guaraní	-55.17	-25.25	South America	1360	2	2	NA	21
acm	Achumawi	Palaihnihan	-121	41.5	North America	16	5	2	2	23
aco	Acoma	Keresan	-107.58	34.92	North America	3391	2	2	3	51
adz	Adzera	Oceanic	146.25	-6.25	SE Asia & Oceania	28900	4	1	NA	25
agh	Aghem	Bantoid	10	6.67	Africa	26727	4	3	2	35
aht	Ahtna	Athapaskan	-145	62	North America	80	1	2	1	35
aik	Aikaná	Arawakan	-60.67	-12.67	South America	90	3	2	2	32
ain	Ainu	Ainu	143	43	Eurasia	15	3	2	2	16
aiz	Aizi	Kru	-4.5	5.25	Africa	6500	3	3	NA	33
akn	Akan	Kwa	-1.25	6.5	Africa	8300000	1	3	2	35
akw	Akawaio	Cariban	-59.5	6	South America	5000	1	3	1	23
abm	Alabama	Muskogean	-87.42	32.33	North America	100	2	1	NA	17
ala	Alamblak	Sepik Hill	143.33	-4.67	Australia-New Guinea	1527	3	3	1	25
alw	Alawa	Maran	134.25	-15.17	Australia-New Guinea	17	4	1	1	26
alb	Albanian	Albanian	20	41	Eurasia	5823075	3	3	1	35
aea	Aleut (Eastern)	Eskimo-Aleut	-164	54.75	North America	490	4	1	1	27
ald	Alladian	Kwa	-4.33	5.17	Africa	23000	1	3	2	36
amc	Amahuaca	Panoan	-72.5	-10.5	South America	110	2	1	NA	22
ame	Amele	Madang	145.58	-5.25	Australia-New Guinea	5300	5	2	1	20
amh	Amharic	Semitic	38	10	Africa	17417913	4	3	1	37
amo	Amo	Kainji	8.67	10.33	Africa	12263	3	3	2	35
amu	Amuesha	Arawakan	-75.42	-10.5	South America	9831	4	1	1	26
amz	Amuzgo	Amuzgoan	-98	16.83	North America	23000	2	3	3	37
adk	Andoke	Andoke	-72	-0.67	South America	619	1	3	2	26
ant	Angaatiha	Angan	146.25	-7.22	Australia-New Guinea	2100	4	2	2	21
anc	Angas	West Chadic	9.5	9.5	Africa	40000	1	3	3	43
ani	//Ani	Central Khoisan	21.92	-18.92	Africa	1000	2	2	2	NA
ao	Ao	Kuki-Chin-Naga	94.67	26.58	SE Asia & Oceania	141000	1	2	3	20
api	Apinayé	Ge-Kaingang	-48	-5.5	South America	800	1	3	1	30
apu	Apurinã	Arawakan	-67	-9	South America	2000	1	2	1	NA
arb	Arabela	Zaparoan	-75.17	-2	South America	50	4	2	1	18
aeg	Arabic (Egyptian)	Semitic	31	30	Africa	46321000	3	2	1	35
ana	Araona	Tacanan	-67.75	-12.33	South America	81	5	1	1	NA
arp	Arapesh	Kombio-Arapesh	143.17	-3.47	Australia-New Guinea	20865	2	3	1	NA
arc	Archi	Lezgic	46.83	42	Eurasia	1000	4	2	1	91

arm	Armenian (Eastern)	Armenian	45	40	Eurasia	6723840	1	2	1	36
amp	Arrernte (Mparntwe)	Pama-Nyungan	136	-24	Australia-New Guinea	2175	4	1	1	30
asm	Asmat	Asmat-Kamoro	138.5	-5.5	Australia-New Guinea	290	3	2	1	17
ata	Atayal	Atayalic	121.33	24.5	SE Asia & Oceania	84330	1	2	1	26
ava	Avar	Avar-Andic-Tsezic	46.5	42.5	Eurasia	600959	5	2	1	49
awp	Awa Pit	Barbacoan	-78.25	1.5	South America	21000	1	1	1	NA
awn	Awngi	Central Cushitic	36.67	10.83	Africa	356980	4	3	2	35
aym	Aymara	Aymaran	-69	-17	South America	2227642	4	1	1	NA
aze	Azerbaijani	Turkic	48.5	40.5	Eurasia	31423529	3	3	1	33
bag	Bagirmi	Bongo-Bagirmi	16	11.67	Africa	44761	4	3	3	NA
bai	Bai	Bai	100	26	SE Asia & Oceania	8e+05	2	3	3	29
bng	Baining	Baining-Taulil	152	-4.58	Australia-New Guinea	6350	1	2	1	22
baj	Bajau	Sama-Bajaw	123	-4.33	SE Asia & Oceania	90000	4	2	1	32
bki	Bakairí	Cariban	-55	-14	South America	570	2	3	1	29
bam	Bambara	Western Mande	-7.5	12.5	Africa	2786385	3	3	2	35
byu	Bandjalang	Pama-Nyungan	153	-27.92	Australia-New Guinea	10	1	1	1	16
bno	Barasano (Northern)	Tucanoan	-70.25	0.33	South America	700	1	2	2	23
brd	Bardi	Nyulnyulan	122.92	-16.58	Australia-New Guinea	20	2	1	1	24
brb	Bariba	Gur	2.5	10	Africa	560000	2	3	3	30
bsk	Bashkir	Turkic	58	53	Eurasia	1871383	4	3	1	38
bsq	Basque	Basque	-3	43	Eurasia	588108	3	2	1	28
bkr	Batak (Karo)	Sundic	98.25	3.25	SE Asia & Oceania	6e+05	2	3	1	21
bto	Batak (Toba)	Sundic	99	2.5	SE Asia & Oceania	2e+06	2	2	1	NA
baw	Bawm	Kuki-Chin-Naga	92.25	22.5	SE Asia & Oceania	13793	3	2	2	NA
bee	Beembe	Bantoid	14.08	-3.92	Africa	3200	2	2	2	26
bej	Beja	Beja	36	18	Africa	1178000	3	2	2	26
bco	Bella Coola	Bella Coola	-126.67	52.5	North America	20	4	1	1	31
ben	Bengali	Indic	90	24	Eurasia	171070202	4	3	1	43
bma	Berber (Middle Atlas)	Berber	-5	33	Africa	3150000	3	1	1	NA
ber	Berta	Berta	34.67	10.33	Africa	146799	3	2	2	29
bet	Bété	Kru	-6.25	6.25	Africa	130000	3	3	3	37
bir	Birom	Platoid	8.83	9.67	Africa	3e+05	3	3	3	29
bis	Bisa	Eastern Mande	-0.5	11.5	Africa	581900	2	2	1	24
bbf	Bobo Fing	Western Mande	-4.42	11.83	Africa	365091	3	3	3	33
bod	Bodo	Baric	92	26.83	SE Asia & Oceania	603301	2	2	2	21
brr	Bororo	Bororo	-57	-16	South America	850	1	3	1	20
brh	Brahui	Northern Dravidian	67	28.5	Eurasia	2210000	3	2	1	33
bra	Brao	Bahnaric	107.5	14.17	SE Asia & Oceania	12800	3	3	NA	31
bre	Breton	Celtic	-3	48	Eurasia	532722	4	3	1	45
bri	Bribri	Talamanca	-83	9.42	South America	11000	2	2	NA	27
brw	Bru (Western)	Katuic	104.75	16.75	SE Asia & Oceania	20000	3	3	1	42
bul	Bulgarian	Slavic	25	42.5	Eurasia	8954811	5	2	1	42
bua	Burarra	Burarran	134.58	-12.25	Australia-New Guinea	400	2	2	1	21
brm	Burmese	Burmese-Lolo	96	21	SE Asia & Oceania	32301581	4	3	3	46
bur	Burushaski	Burushaski	74.5	36.5	Eurasia	87049	5	2	1	43
cac	Cacua	Cacua-Nukak	-70	1.08	South America	150	1	2	2	22

cad	Caddo	Caddoan	-93.5	33.33	North America	25	3	1	2	23
cah	Cahuilla	Takic	-116.25	33.5	North America	7	3	1	1	NA
cax	Campa (Axininca)	Arawakan	-74	-12	South America	23750	2	1	1	19
cam	Camsá	Camsá	-77	1.17	South America	4022	3	2	1	28
ckr	Canela-Krahô	Ge-Kaingang	-45	-6	South America	2620	1	3	1	NA
cnt	Cantonese	Chinese	113	23	SE Asia & Oceania	54810598	3	3	3	NA
car	Carib	Cariban	-56	5.5	South America	10226	2	2	1	22
ctl	Catalan	Romance	2	41.75	Eurasia	6667328	3	3	1	NA
cay	Cayapa	Barbacoan	-79	0.67	South America	9500	3	1	NA	28
cyv	Cayuvava	Cayuvava	-65.5	-13.5	South America	NA	2	3	1	33
chw	Cham (Western)	Sundic	105.5	12	SE Asia & Oceania	253100	3	3	2	32
cha	Chamorro	Chamorro	144.75	13.45	SE Asia & Oceania	76705	3	2	1	26
cso	Chatino (Sierra Occ.)	Zapotecan	-97.33	16.25	North America	12000	2	2	3	25
chl	Chehalis (Upper)	Tsamosan	-123	46.58	North America	NA	4	1	1	34
che	Cherokee	Southern Iroquoian	-83.5	35.5	North America	15000	1	2	2	17
cck	Chickasaw	Muskogean	-88	34	North America	1000	2	1	1	NA
cti	Chin (Tiddim)	Kuki-Chin-Naga	93.67	23.33	SE Asia & Oceania	344100	3	2	3	52
cle	Chinantec (Lealao)	Chinantecan	-95.92	17.33	North America	2000	3	2	3	NA
chq	Chinantec (Quiotepec)	Chinantecan	-96.67	17.58	North America	8000	4	3	3	41
chp	Chipewyan	Athapaskan	-106	59	North America	4000	5	2	2	52
cve	Chuave	Chimbu	145.12	-6.12	Australia-New Guinea	23100	1	2	NA	17
chk	Chukchi	N. Chukotko-Kamchatkan	-173	67	Eurasia	10000	2	2	1	22
chu	Chulupí	Matacoan	-60.5	-23.5	South America	18200	3	1	1	28
chv	Chuvash	Turkic	47.5	55.5	Eurasia	1834394	3	3	1	30
cil	CiLuba	Bantoid	22	-6	Africa	6300000	3	2	2	NA
ccp	Cocopa	Yuman	-115	32.33	North America	350	3	1	1	NA
cof	Cofán	Cofán	-77.17	0.17	South America	800	4	2	NA	35
cmn	Comanche	Numic	-101.5	33.5	North America	200	1	2	1	NA
coo	Coos (Hanis)	Coosan	-124.17	43.5	North America	1	5	2	1	NA
cre	Cree (Plains)	Algonquian	-110	54	North America	34100	1	1	1	NA
cub	Cubeo	Tucanoan	-70.5	1.33	South America	6150	1	2	2	23
dad	Dadibi	Teberan	144.58	-6.55	Australia-New Guinea	10000	1	2	2	23
dag	Daga	Dagan	149.33	-10	Australia-New Guinea	9000	1	2	NA	NA
dgb	Dagbani	Gur	-0.5	9.58	Africa	8e+05	3	2	2	29
dgr	Dagur	Mongolic	124	48	Eurasia	96085	3	3	1	29
dah	Dahalo	Southern Cushitic	40.5	-2.33	Africa	400	5	2	2	59
ddf	Daju (Dar Fur)	Daju	25.25	12.25	Africa	143053	3	2	1	30
dan	Dan	Eastern Mande	-8	7.5	Africa	951600	3	3	3	39
dnw	Dangaléat (Western)	East Chadic	18.33	12.17	Africa	45000	3	3	2	28
dni	Dani (L. Grand Valley)	Dani	138.83	-4.33	Australia-New Guinea	20000	1	3	1	24
dar	Darai	Indic	84	24	Eurasia	10210	4	2	1	NA
der	Dera	Senagi	141	-3.58	Australia-New Guinea	1687	1	2	1	17
det	Deti	Central Khoisan	24.5	-20.5	Africa	6000	5	2	2	NA
die	Diegueño	Yuman	-116.17	32.67	North America	295	4	2	1	34
din	Dinka	Nilotic	28	8.5	Africa	320000	3	3	3	32
dio	Diola-Fogny	Northern Atlantic	-16	13	Africa	358276	3	3	1	29

diy	Diyari	Pama-Nyungan	139	-28	Australia-New Guinea	NA	3	1	1	25
diz	Dizi	Omotic	36.5	6.17	Africa	21075	3	2	3	30
djp	Djapu	Pama-Nyungan	136	-12.67	Australia-New Guinea	500	3	1	1	23
dts	Dogon (Toro So)	Dogon	-3.33	14.5	Africa	50000	2	3	2	28
doy	Doyayo	Adamawa-Ubangian	13.08	8.67	Africa	18000	3	3	3	34
dre	Drehu	Oceanic	167.25	-21	SE Asia & Oceania	11338	4	3	1	NA
dum	Dumo	Western Sko	141.3	-2.67	Australia-New Guinea	2667	1	3	3	28
dyi	Dyirbal	Pama-Nyungan	145.58	-17.83	Australia-New Guinea	40	1	1	1	16
efi	Efik	Cross River	8.5	4.92	Africa	4e+05	1	3	2	20
eja	Ejagham	Bantoid	8.67	5.42	Africa	116675	3	3	3	27
eka	Ekari	Wissel Lakes-Kemandoga	135.5	-3.83	Australia-New Guinea	1e+05	1	2	2	15
eng	English	Germanic	0	52	Eurasia	309582484	3	3	1	NA
epe	Epena Pedee	Choco	-77	3	South America	8050	2	3	1	31
evn	Even	Tungusic	130	68	Eurasia	7543	2	2	1	27
eve	Evenki	Tungusic	125	56	Eurasia	29000	2	2	1	NA
ewe	Ewe	Kwa	0.42	6.33	Africa	3112400	4	3	2	40
ewo	Ewondo	Bantoid	12	4	Africa	577700	4	3	3	34
eya	Eyak	Eyak	-145	60.5	North America	1	4	1	1	45
fas	Fasu	Kutubuan	143.33	-6.58	Australia-New Guinea	1200	1	2	2	21
fef	Fefe	Bantoid	10.17	5.25	Africa	123700	2	3	3	25
fij	Fijian	Oceanic	178	-17.83	SE Asia & Oceania	334061	3	2	1	30
fin	Finnish	Finnic	25	62	Eurasia	5232728	2	3	1	25
fre	French	Romance	2	48	Eurasia	64858311	3	3	1	37
ful	Fulniô	Yatê	-37.5	-8	South America	2788	3	3	2	30
fur	Fur	Fur	25	13.5	Africa	501800	2	2	2	30
fuz	Fuzhou	Chinese	119.5	26	SE Asia & Oceania	9103157	1	3	3	21
fye	Fyem	Platoid	9.33	9.58	Africa	3000	5	2	2	NA
ga	Gã	Kwa	-0.17	5.67	Africa	6e+05	4	3	3	41
gds	Gadsup	Eastern Highlands	146	-6.25	Australia-New Guinea	22061	1	1	3	15
gar	Garo	Baric	90.5	25.67	SE Asia & Oceania	677000	2	2	1	NA
grr	Garrwa	Garrwan	137.17	-17.08	Australia-New Guinea	200	3	1	1	22
gbb	Gbeya Bossangoa	Adamawa-Ubangian	17.5	6.67	Africa	176000	4	3	2	43
gla	Gelao	Kadai	105.5	22.92	SE Asia & Oceania	3000	4	3	3	43
geo	Georgian	Kartvelian	44	42	Eurasia	4178604	4	2	1	34
ger	German	Germanic	10	52	Eurasia	95392978	3	3	1	41
goa	Goajiro	Arawakan	-72	12	South America	135000	1	2	NA	26
goo	Gooniyandi	Bunuban	126.33	-18.33	Australia-New Guinea	100	3	1	1	NA
gan	Great Andamanese	Great Andamanese	92.67	12	SE Asia & Oceania	24	1	3	1	24
grb	Grebo	Kru	-8	5	Africa	23700	3	3	3	NA
grk	Greek (Modern)	Greek	22	39	Eurasia	12258540	3	2	1	26
grw	Greenlandic (West)	Eskimo-Aleut	-51	64	North America	54800	3	1	1	22
ghb	Guahibo	Guahiban	-69	5	South America	23000	2	2	NA	29
gmb	Guambiano	Barbacoan	-76.67	2.5	South America	23500	2	2	NA	24
gua	Guaraní	Tupi-Guaraní	-56	-26	South America	4848000	3	2	1	36
gwa	Gwari	Nupoid	7	9.5	Africa	1050000	3	2	3	26
had	Hadza	Hadza	35.17	-3.75	Africa	800	5	2	2	62

hai	Haida	Haida	-132	53	North America	55	5	1	1	49
hak	Hakka	Chinese	116	25	SE Asia & Oceania	29937959	2	2	3	22
hmr	Hamer	Omotic	36.5	5	Africa	42838	4	2	2	35
ham	Hamtai	Angan	146.25	-7.5	Australia-New Guinea	45000	1	3	2	NA
hau	Hausa	West Chadic	7	12	Africa	24162000	4	2	2	38
haw	Hawaiian	Oceanic	-155.5	19.58	SE Asia & Oceania	1000	1	2	1	13
hba	Hebrew (Modern)	Semitic	35.17	31.75	Africa	5055000	2	2	1	NA
hin	Hindi	Indic	77	25	Eurasia	180764791	5	2	1	61
hix	Hixkaryana	Cariban	-59	-1	South America	600	2	2	1	23
hmo	Hmong Njua	Hmong-Mien	105	28	SE Asia & Oceania	1290600	5	2	3	56
hop	Hopi	Hopi	-110	36	North America	5264	3	2	2	28
htc	Huastec	Mayan	-99.33	22.08	North America	1749	3	2	1	26
hve	Huave (Mateo d. Mar)	Huavean	-95	16.22	North America	12000	3	2	2	29
hum	Huitoto (Murui)	Huitoto	-73.5	-1	South America	2900	2	2	1	NA
hun	Hungarian	Ugric	20	47	Eurasia	13611600	4	3	1	40
hzb	Hunzib	Avar-Andic-Tsezic	46.25	42.17	Eurasia	2000	4	3	1	NA
hup	Hupa	Athapaskan	-123.67	41.08	North America	8	4	1	1	35
iaa	Iaai	Oceanic	166.58	-20.42	SE Asia & Oceania	1562	5	3	1	52
iba	Iban	Sundic	112	2	SE Asia & Oceania	415000	3	2	1	25
igb	Igbo	Igboid	7.33	6	Africa	1.8e+07	5	3	2	59
ign	Ignaciano	Arawakan	-65.42	-15.17	South America	4500	3	1	NA	25
ijo	Ijo (Kolokuma)	Ijoid	5.67	4.92	Africa	1e+06	3	3	2	37
ik	Ik	Kuliak	34.17	3.75	Africa	2000	4	3	2	44
ika	Ika	Aruak	-73.75	10.67	South America	14301	2	3	1	NA
imo	Imonda	Border	141.17	-3.33	Australia-New Guinea	250	1	3	1	NA
ind	Indonesian	Sundic	106	0	SE Asia & Oceania	23143354	3	2	1	NA
igs	Ingessana	Eastern Jebel	34	11.5	Africa	67166	2	2	2	33
ing	Ingush	Nakh	45.08	43.17	Eurasia	230315	5	2	1	NA
irx	Iranxe	Arawakan	-58	-13	South America	191	3	2	1	39
irq	Iraqw	Southern Cushitic	35.5	-4	Africa	462000	4	2	2	45
irr	Irarutu	South Halmahera (WNG)	133.5	-3	SE Asia & Oceania	4000	1	3	1	19
ird	Irish (Donegal)	Celtic	-8	55	Eurasia	355000	5	2	1	69
iso	Isoko	Edoid	6.25	5.5	Africa	423000	4	3	2	37
ite	Itelmen	S. Chukotko-Kamchatkan	157.5	57	Eurasia	380	4	2	NA	32
ito	Itonama	Itonama	-64.33	-12.83	South America	10	3	2	NA	25
ivs	Ivatan (Southern)	Northern Philippines	121.83	20.33	SE Asia & Oceania	35000	3	1	1	23
iwm	Iwam	Upper Sepik	142	-4.33	Australia-New Guinea	3000	1	2	NA	17
jak	Jakaltek	Mayan	-91.67	15.67	North America	99000	4	2	1	32
jpn	Japanese	Japanese	140	37	Eurasia	122433899	2	2	2	20
jpr	Japreria	Cariban	-73	10.5	South America	90	1	2	1	24
jaq	Jaqaru	Aymaran	-76	-13	South America	736	5	1	1	39
jav	Javanese	Sundic	111	-7	SE Asia & Oceania	75508300	3	3	1	29
jeb	Jebero	Cahuapanan	-76.5	-5.42	South America	2500	3	1	1	23
jeh	Jeh	Bahnaric	107.83	15.17	SE Asia & Oceania	23256	5	3	1	NA
jng	Jingpho	Jinghpo	97	25.42	SE Asia & Oceania	940000	4	2	3	30
jiv	Jivaro	Jivaroan	-78	-2.5	South America	46700	2	1	NA	23

jom	Jomang	Kordofanian	30.5	10.58	Africa	1500	1	3	2	21
kek	Kekchí	Mayan	-89.83	16	North America	4e+05	3	2	NA	26
kab	Kabardian	Northwest Caucasian	43.5	43.5	Eurasia	1012000	5	1	1	56
kad	Kadugli	Kadugli	29.67	11	Africa	81500	3	2	3	27
kng	Kaingang	Ge-Kaingang	-52	-26	South America	18000	1	3	1	27
kly	Kala Lagaw Ya	Pama-Nyungan	142.12	-10.12	Australia-New Guinea	3000	2	2	1	24
kal	Kalami	Indic	72.5	35.5	Eurasia	40000	2	2	2	NA
kkv	Kaliai-Kove	Oceanic	149.67	-5.58	SE Asia & Oceania	8750	3	2	1	21
kgu	Kalkatungu	Pama-Nyungan	139.5	-21	Australia-New Guinea	NA	3	1	1	23
kzh	Kam (Zhanglu)	Kam-Tai	108.5	26	SE Asia & Oceania	463000	3	3	3	27
knk	Kanakuru	West Chadic	12	10	Africa	20000	4	2	2	35
knd	Kannada	Southern Dravidian	76	14	Eurasia	35346000	3	2	1	NA
knr	Kanuri	Saharan	13	12	Africa	3425138	4	3	2	29
ksg	Karen (Sgaw)	Karen	97	18	SE Asia & Oceania	1584700	2	3	3	36
krk	Karok	Karok	-123	41.67	North America	10	4	2	2	27
kas	Kashmiri	Indic	76	34	Eurasia	4611000	4	3	1	55
kws	Kawaiisu	Numic	-117.5	36	North America	8	3	2	1	31
kyl	Kayah Li (Eastern)	Karen	97.5	19	SE Asia & Oceania	360220	2	3	3	NA
kay	Kayardild	Tangkic	139.5	-17.05	Australia-New Guinea	6	2	1	1	NA
ked	Kedang	Central Malayo-Polynesian	123.75	-8.25	SE Asia & Oceania	30000	3	2	1	NA
kef	Kefa	Omotic	36.25	7.25	Africa	569626	3	2	2	27
ker	Kera	East Chadic	15.08	9.83	Africa	50523	2	2	3	30
ket	Ket	Yeniseian	87	64	Eurasia	550	2	3	1	25
kew	Kewa	Engan	143.83	-6.5	Australia-New Guinea	90000	3	2	2	20
kha	Khalkha	Mongolic	105	47	Eurasia	2337095	3	3	1	33
kty	Khanty	Ugric	65	65	Eurasia	12000	2	2	1	32
khr	Kharia	Munda	84.33	22.5	Eurasia	293575	4	2	1	36
khs	Khasi	Khasian	92	25.5	SE Asia & Oceania	865000	3	2	1	29
khm	Khmer	Khmer	105	12.5	SE Asia & Oceania	13276639	3	3	1	42
kmu	Khmu	Palaung-Khmuic	102	21	SE Asia & Oceania	479739	3	3	2	41
kho	Khoekhoe	Central Khoisan	18	-25.5	Africa	233701	4	2	3	41
klv	Kilivila	Oceanic	151.08	-8.5	SE Asia & Oceania	20000	3	2	1	NA
kio	Kiowa	Kiowa-Tanoan	-99	37	North America	1092	3	2	2	42
kgz	Kirghiz	Turkic	75	42	Eurasia	3136733	3	3	1	30
krb	Kiribati	Oceanic	173	1.33	SE Asia & Oceania	67790	1	2	1	NA
kss	Kisi (Southern)	Southern Atlantic	-10.25	8.5	Africa	2e+05	2	3	2	NA
kiw	Kiwai	Kiwaian	143.5	-8	Australia-New Guinea	14100	1	2	2	19
klm	Klamath	Klamath-Modoc	-121.5	42.5	North America	1	4	1	1	37
kla	Klao	Kru	-8.75	4.75	Africa	192000	1	3	3	27
koa	Koasati	Muskogean	-85.17	34.83	North America	200	1	1	3	NA
kob	Kobon	Madang	144.33	-5.17	Australia-New Guinea	6000	3	3	1	NA
koh	Kohumono	Cross River	8.12	6	Africa	30000	4	3	3	38
koi	Koiari	Koiarian	147.33	-9.5	Australia-New Guinea	1700	1	2	2	16
kzy	Komi-Zyrian	Finnic	55	65	Eurasia	262200	4	3	1	33
kom	Komo	Koman	33.75	8.75	Africa	11500	3	3	3	31
kkn	Konkani	Indic	74	15.25	Eurasia	4e+06	3	3	1	37

kgi	Konyagi	Northern Atlantic	-13.25	12.5	Africa	18400	5	NA	2	46
kor	Korean	Korean	128	37.5	Eurasia	67019690	3	3	1	32
kfe	Koromfe	Gur	-0.92	14.25	Africa	196100	2	3	1	NA
kry	Koryak	N. Chukotko-Kamchatkan	167	61	Eurasia	3500	2	2	1	21
kot	Kota	Southern Dravidian	77.17	11.5	Eurasia	2000	3	2	1	28
ktk	Kotoko	Biu-Mandara	15.33	11.33	Africa	30000	4	3	2	36
koy	Koya	South-Central Dravidian	81.33	17.5	Eurasia	330000	3	2	1	24
kch	Koyra Chiini	Songhay	-3	17	Africa	2e+05	2	2	1	NA
kse	Koyraboro Senni	Songhay	0	16	Africa	1e+05	2	2	NA	24
kpa	Kpan	Platoid	10.17	7.58	Africa	11386	3	2	3	34
kpe	Kpelle	Western Mande	-10	7	Africa	487400	3	3	3	34
kro	Krongo	Kadugli	30	10.5	Africa	21688	3	3	2	NA
kya	Kuku-Yalanji	Pama-Nyungan	145	-16	Australia-New Guinea	700	1	1	1	16
kul	Kullo	Omotic	37.08	6.75	Africa	1236637	3	2	2	29
kun	Kuna	Kuna	-77.33	8	South America	1576	2	2	1	21
knm	Kunama	Kunama	37	14.5	Africa	108883	3	2	2	26
kmp	Kunimaipa	Goilalan	146.83	-8	Australia-New Guinea	11000	2	2	1	20
krd	Kurdish (Central)	Iranian	44	36	Eurasia	9113505	4	3	1	47
kur	Kurukh	Northern Dravidian	85.5	22.83	Eurasia	2050000	3	2	NA	32
kut	Kutenai	Kutenai	-116	49.5	North America	12	4	1	1	NA
kwa	Kwaio	Oceanic	161	-8.95	SE Asia & Oceania	13249	2	2	1	21
kwk	Kwakwala	Northern Wakashan	-127	51	North America	235	5	2	1	48
kwo	Kwoma	Middle Sepik	142.75	-4.17	Australia-New Guinea	3000	3	3	NA	31
lad	Ladakhi	Bodic	78	34	SE Asia & Oceania	114000	4	2	1	NA
lah	Lahu	Burmese-Lolo	98.17	20	SE Asia & Oceania	577178	4	3	3	35
lak	Lak	Lak-Dargwa	47.17	42.17	Eurasia	119512	5	1	1	69
lkt	Lakhota	Siouan	-101.83	43.83	North America	6000	4	2	1	36
lkk	Lakkia	Kadai	110.17	24.08	SE Asia & Oceania	12000	5	2	3	55
lam	Lamé	Masa	14.5	9	Africa	35720	4	2	3	38
lan	Lango	Nilotic	33	2.17	Africa	977680	4	3	2	NA
lat	Latvian	Baltic	24	57	Eurasia	1543844	2	2	2	NA
lav	Lavukaleve	Solomons East Papuan	159.2	-9.08	Australia-New Guinea	1783	4	2	1	NA
llm	Lelemi	Kwa	0.5	7.33	Africa	48900	3	3	2	34
len	Lenakel	Oceanic	169.25	-19.45	SE Asia & Oceania	6500	2	2	1	21
lep	Lepcha	Lepcha	88.5	27.17	SE Asia & Oceania	48000	4	3	NA	NA
lez	Lezgian	Lezgic	47.83	41.67	Eurasia	451112	5	2	1	NA
lit	Lithuanian	Baltic	24	55	Eurasia	2960000	5	2	NA	52
lu	Lü	Kam-Tai	100.67	22	SE Asia & Oceania	672064	3	3	3	31
lua	Lua	Adamawa-Ubangian	17.75	9.75	Africa	5157	3	3	3	36
lug	Lugbara	Moru-Ma'di	30.92	3.08	Africa	1040000	4	3	3	36
lui	Luiseño	Takic	-117.17	33.33	North America	30	3	2	1	26
luo	Luo	Nilotic	34.75	-0.5	Africa	3465000	3	3	2	32
lus	Lushootseed	Central Salish	-122	48	North America	60	5	1	1	37
luv	Luvale	Bantoid	22	-12	Africa	669000	2	2	2	NA
mya	Maya	South Halmahera (WNG)	130.92	-1.25	SE Asia & Oceania	4000	1	2	3	NA
maa	Maasai	Nilotic	36	-3	Africa	883000	3	3	2	28

mab	Maba	Maban	20.83	13.75	Africa	250000	3	3	2	29
mne	Maidu (Northeast)	Maiduan	-120.67	40	North America	1	2	2	1	23
mal	Malagasy	Borneo	47	-20	SE Asia & Oceania	5948700	3	1	1	25
mlk	Malakmalak	Northern Daly	130.42	-13.42	Australia-New Guinea	9	1	2	1	19
mla	Mambila	Bantoid	11.5	6.75	Africa	129000	3	3	3	25
mnc	Manchu	Tungusic	127.5	49.5	Eurasia	60	2	2	NA	25
mnd	Mandarin	Chinese	110	34	SE Asia & Oceania	873014298	4	2	3	32
myi	Mangarrayi	Mangarrayi	133.5	-14.67	Australia-New Guinea	50	2	2	1	NA
mgg	Mangghuer	Mongolic	102	36	Eurasia	152000	3	2	1	28
mao	Maori	Oceanic	176	-40	SE Asia & Oceania	50000	1	2	1	NA
map	Mapudungun	Araucanian	-72	-38	South America	3e+05	3	2	1	26
mrn	Maranao	Southern Philippines	124.25	7.83	SE Asia & Oceania	776169	1	1	1	17
mku	Maranungku	Western Daly	130	-13.67	Australia-New Guinea	15	1	2	1	NA
mrg	Margi	Biu-Mandara	13	11	Africa	158000	4	1	2	34
mme	Mari (Meadow)	Finnic	48	57	Eurasia	451000	4	3	NA	33
mar	Maricopa	Yuman	-113.17	33.17	North America	181	3	2	1	NA
mrd	Marind	Marind Proper	140.17	-7.83	Australia-New Guinea	7000	2	2	1	NA
mrt	Martuthunira	Pama-Nyungan	116.5	-20.83	Australia-New Guinea	5	3	1	1	NA
mau	Maung	Iwaidjan	133.5	-11.92	Australia-New Guinea	200	2	2	1	22
max	Maxakalí	Maxakalí	-40	-18	South America	728	1	2	1	20
may	Maybrat	North-Central Bird's Head	132.5	-1.33	Australia-New Guinea	20000	1	2	1	NA
maz	Mazahua	Otomian	-99.92	19.42	North America	365000	5	1	2	60
mzc	Mazatec Chiquihuitlán	Popolocan	-96.92	17.75	North America	2500	3	2	3	33
mba	Mba	Adamawa-Ubangian	25	1	Africa	36087	3	3	3	31
mbb	Mbabaram	Pama-Nyungan	145	-17.17	Australia-New Guinea	2	3	1	1	24
mbm	Mbum	Adamawa-Ubangian	13.17	7.75	Africa	38600	4	NA	2	38
mei	Meithei	Kuki-Chin-Naga	94	24.75	SE Asia & Oceania	1261000	4	2	2	NA
mie	Mien	Hmong-Mien	111	25	SE Asia & Oceania	818685	4	3	3	41
mss	Miwok (S. Sierra)	Miwok	-120	37.5	North America	7	2	2	1	21
mtp	Mixe (Totontepec)	Mixe-Zoque	-96	17.25	North America	5200	1	3	1	23
mxc	Mixtec (Chalcatongo)	Mixtecan	-97.58	17.05	North America	14453	2	2	3	25
mxm	Mixtec (Molinos)	Mixtecan	-97.58	17	North America	14453	2	2	3	NA
mog	Moghol	Mongolic	62	35	Eurasia	200	3	2	1	29
mor	Mor	South Halmahera (WNG)	135.75	-3	SE Asia & Oceania	700	1	2	2	19
mro	Moro	Kordofanian	30.17	11	Africa	30000	3	3	2	29
mov	Movima	Movima	-65.67	-13.83	South America	1452	2	2	1	23
mui	Muinane	Boran	-72.5	-1	South America	150	3	2	2	28
mum	Mumuye	Adamawa-Ubangian	11.67	9	Africa	4e+05	3	2	3	34
mun	Mundari	Munda	84.67	23	Eurasia	2074700	4	2	1	37
mrl	Murle	Surmic	33.5	6.5	Africa	60200	3	3	2	26
mpa	Murrinh-Patha	Murrinh-Patha	129.67	-14.67	Australia-New Guinea	900	3	1	1	25
nhn	Nahuatl (N. Puebla)	Aztecan	-98.25	20	North America	60000	2	1	1	20
nht	Nahuatl (Tetelcingo)	Aztecan	-99	19.67	North America	3500	2	2	1	NA
nbk	Nambakaengö	Reef Islands - Santa Cruz	165.87	-10.78	Australia-New Guinea	4280	4	3	NA	47
nmb	Nambikuára	Nambikuaran	-59	-13	South America	1150	4	2	3	43
nai	Nanai	Tungusic	137	49.5	Eurasia	5772	2	2	1	24

nnc	Nancowry	Nicobarese	93.5	8.05	SE Asia & Oceania	2200	2	3	1	25
nan	Nandi	Nilotic	35	0.25	Africa	2458123	1	3	2	NA
nar	Nara (in Ethiopia)	Nara	37.58	15.08	Africa	80000	2	2	2	22
nas	Nasioi	East Bougainville	155.58	-6.33	Australia-New Guinea	20000	1	2	NA	13
nav	Navajo	Athapaskan	-108	36.17	North America	148530	4	1	2	47
nax	Naxi	Naxi	100	27.5	SE Asia & Oceania	308839	5	3	3	49
ndt	Ndut	Northern Atlantic	-16.92	14.92	Africa	35000	4	3	3	34
ndy	Ndyuka	Creoles and Pidgins	-54.5	5	South America	15500	2	2	2	NA
nen	Nenets	Samoyedic	72	69	Eurasia	26730	4	2	1	35
nap	Neo-Aramaic	Semitic	47	38	Africa	4378	3	2	1	40
nep	Nepali	Indic	85	28	Eurasia	17209255	4	2	1	39
new	Newari (Kathmandu)	Bodic	85.5	27.67	SE Asia & Oceania	825458	3	1	1	32
nez	Nez Perce	Sahaptian	-116	46	North America	100	4	2	1	30
nga	Nganasan	Samoyedic	93	71	Eurasia	500	2	3	1	29
nti	Ngiti	Lendu	30.25	1.33	Africa	1e+05	5	3	3	NA
ngz	Ngizim	West Chadic	10.92	12.08	Africa	80000	4	2	2	40
nim	Nimboran	Nimboran	140.17	-2.5	Australia-New Guinea	2000	1	2	NA	18
nis	Nishi	Mirish	93.5	27.5	SE Asia & Oceania	261000	2	3	2	24
niv	Nivkh	Nivkh	142	53.33	Eurasia	1089	4	2	2	35
nko	Nkore-Kiga	Bantoid	29.83	-0.92	Africa	1391442	3	2	2	NA
nob	Nobiin	Nubian	31	21	Africa	495000	2	2	2	21
non	Noni	Bantoid	10.58	6.42	Africa	25000	3	3	3	33
nor	Norwegian	Germanic	8	61	Eurasia	4640000	3	3	2	46
nun	Nung (in Vietnam)	Kam-Tai	106.42	21.92	SE Asia & Oceania	856412	3	2	3	32
nug	Nunggubuyu	Nunggubuyu	135.67	-13.75	Australia-New Guinea	300	3	1	1	23
nuu	Nuuchahnulth	Southern Wakashan	-126.67	49.67	North America	200	5	1	1	42
nkt	Nyah Kur (Tha Pong)	Monic	101.67	15.67	SE Asia & Oceania	10000	4	3	2	50
nyg	Nyangi	Kuliak	33.58	3.42	Africa	NA	2	3	2	25
nyi	Nyimang	Nyimang	29.33	12.17	Africa	70000	2	3	3	25
ood	Oodham	Tepiman	-112	32	North America	11819	3	2	1	24
oca	Ocaina	Huitoto	-71.75	-2.75	South America	66	4	2	2	34
ogb	Ogbia	Cross River	6.25	4.67	Africa	2e+05	3	3	2	34
oji	Ojibwa (Eastern)	Algonquian	-80	46	North America	25885	2	1	1	27
ond	Oneida	Northern Iroquoian	-75.67	43	North America	250	4	1	2	NA
orm	Ormuri	Iranian	69.75	32.5	Eurasia	1050	1	2	1	31
orh	Oromo (Harar)	Eastern Cushitic	42	9	Africa	4526000	4	2	3	NA
otm	Otomí (Mezquital)	Otomian	-99.17	20.17	North America	1e+05	3	3	2	NA
pms	Paamese	Oceanic	168.25	-16.5	SE Asia & Oceania	6000	2	2	1	NA
pac	Pacoh	Katuic	107.08	16.42	SE Asia & Oceania	29224	2	3	1	33
pae	Páez	Páezan	-76	2.67	South America	71400	5	1	1	37
pai	Paiwan	Paiwanic	120.83	22.5	SE Asia & Oceania	66084	3	1	1	26
pnr	Panare	Cariban	-66	6.5	South America	1200	1	3	1	25
puk	Parauk	Palaung-Khmuic	99.5	23.25	SE Asia & Oceania	528400	4	3	1	77
psh	Pashto	Iranian	67	33	Eurasia	7922657	4	2	1	38
psm	Passamaquoddy-M.	Algonquian	-67	45	North America	1655	2	2	2	NA
pau	Paumarí	Arauan	-64	-6	South America	700	3	1	1	NA

paw	Pawaian	Pawaian	145.08	-7	Australia-New Guinea	4000	2	2	2	NA
pec	Pech	Paya	-85.5	15	South America	994	4	2	2	28
prs	Persian	Iranian	54	32	Eurasia	24316121	3	2	1	30
phl	Phlong	Karen	99	15	SE Asia & Oceania	60000	1	2	3	37
prh	Pirahã	Mura	-62	-7	South America	150	2	1	2	11
pit	Pitjantjatjara	Pama-Nyungan	130	-26	Australia-New Guinea	2500	3	1	1	NA
poa	Po-Ai	Oceanic	164.83	-20.67	SE Asia & Oceania	1131	2	3	3	35
poh	Pohnpeian	Oceanic	158.25	6.88	SE Asia & Oceania	29000	1	3	1	20
pol	Polish	Slavic	20	52	Eurasia	42708133	4	2	1	NA
pso	Pomo (Southeastern)	Pomoan	-122.5	39	North America	5	2	2	NA	32
pur	Purépecha	Tarascan	-101.67	19.5	North America	120000	4	2	1	39
qaw	Qawasqar	Alacalufan	-75	-49	South America	20	4	1	1	19
qco	Quechua (Cochab.)	Quechuan	-66	-17.5	South America	3637500	4	2	1	36
qui	Quileute	Chimakuan	-124.25	47.92	North America	10	4	1	2	37
ram	Rama	Rama	-83.75	11.75	South America	24	2	1	1	NA
rap	Rapanui	Oceanic	-109	-27	SE Asia & Oceania	3392	1	2	1	NA
res	Resígaro	Arawakan	-71.5	-2.42	South America	14	4	2	2	35
rom	Romanian	Romance	25	46	Eurasia	23498367	3	3	1	32
rsc	Romansch (Scharans)	Romance	9.5	46.75	Eurasia	40000	3	2	1	NA
ror	Roro	Oceanic	146.58	-8.75	SE Asia & Oceania	15000	1	2	1	14
rtk	Rotokas	West Bougainville	155.17	-6	Australia-New Guinea	4320	1	2	1	11
ruk	Rukai	Tsouic	120.83	22.83	SE Asia & Oceania	10543	3	1	1	27
rus	Russian	Slavic	38	56	Eurasia	145031551	4	2	1	38
rut	Rutul	Lezgic	47.42	41.5	Eurasia	20111	5	2	2	64
sab	Saban	Borneo	115.67	3.67	SE Asia & Oceania	1110	3	3	NA	26
scs	Saami (Central-South)	Finnic	16.75	64.67	Eurasia	600	5	2	1	45
sba	Sáliba (in Colombia)	Sáliban	-70	6	South America	1555	3	2	1	32
sdw	Sandawe	Sandawe	35	-5	Africa	40000	5	2	2	54
san	Sango	Adamawa-Ubangian	18	5	Africa	404000	4	3	2	37
snm	Sanuma	Yanomam	-64.67	4.5	South America	5074	1	3	1	NA
svs	Savosavo	Solomons East Papuan	159.8	-9.13	Australia-New Guinea	2415	2	2	1	22
seb	Sebei	Nilotic	34.58	1.33	Africa	181000	1	2	NA	26
sed	Sedang	Bahnaric	108	14.83	SE Asia & Oceania	101434	5	3	1	55
slp	Selepet	Finisterre-Huon	147.17	-6.17	Australia-New Guinea	7000	2	2	NA	21
sel	Selknam	Chon Proper	-70	-53	South America	1	3	1	1	NA
skp	Selkup	Samoyedic	82	65	Eurasia	1640	2	3	NA	34
sml	Semelai	Aslian	103	3	SE Asia & Oceania	2932	4	3	1	NA
snd	Senadi	Gur	-6.25	9.5	Africa	862000	3	3	3	36
snc	Seneca	Northern Iroquoian	-77.5	42.5	North America	175	1	2	1	19
snt	Sentani	Sentani	140.58	-2.58	Australia-New Guinea	30000	1	3	NA	17
sha	Shan	Kam-Tai	98	22	SE Asia & Oceania	3260000	3	2	3	25
shs	Shasta	Shasta	-122.67	41.83	North America	NA	2	1	2	21
shk	Shipibo-Konibo	Panoan	-75	-7.5	South America	26000	2	1	1	NA
shi	Shiriana	Yanomam	-62.83	3.5	South America	566	1	2	1	25
shu	Shuswap	Interior Salish	-120	52	North America	500	5	2	1	44
sdh	Sindhi	Indic	69	26	Eurasia	21362000	5	3	1	NA

snh	Sinhala	Indic	80.5	7	Eurasia	13220256	3	3	1	36
sin	Siona	Tucanoan	-76.25	0.33	South America	300	2	2	NA	30
srn	Sirionó	Tupi-Guaraní	-64	-15.58	South America	399	3	3	1	28
sla	Slave	Athapaskan	-125	67	North America	2200	4	2	2	NA
som	Somali	Eastern Cushitic	45	3	Africa	12653480	3	3	2	32
soq	Soqotri	Semitic	54	12.5	Africa	64000	4	2	1	34
sor	Sora	Munda	84.33	20	Eurasia	288000	3	2	1	NA
spa	Spanish	Romance	-4	40	Eurasia	322299171	4	2	1	25
squ	Squamish	Central Salish	-123.17	49.67	North America	15	4	1	1	NA
sre	Sre	Bahnaric	108	11.5	SE Asia & Oceania	128723	1	3	2	37
sue	Suena	Binanderean	147.55	-7.75	Australia-New Guinea	3000	5	2	2	18
sui	Sui	Kam-Tai	107.5	26	SE Asia & Oceania	200120	3	3	3	54
sup	Supyire	Gur	-5.58	11.5	Africa	364000	2	3	3	NA
swa	Swahili	Bantoid	39	-6.5	Africa	772642	4	2	1	NA
tab	Taba	South Halmahera (WNG)	127.5	0	SE Asia & Oceania	20000	3	2	1	NA
tac	Tacana	Tacanan	-68	-13.5	South America	1821	2	1	1	22
tag	Tagalog	Meso-Philippine	121	15	SE Asia & Oceania	15900098	3	2	1	23
tma	Tama	Taman	22	14.5	Africa	62931	3	3	3	30
tam	Tamang	Bodic	85.25	28	SE Asia & Oceania	777234	4	2	2	29
tmp	Tampulma	Gur	-0.58	10.42	Africa	16000	1	3	2	33
tok	Tarok	Platoid	10.08	9	Africa	3e+05	4	2	3	32
tsg	Tausug	Meso-Philippine	121	6	SE Asia & Oceania	1022000	4	1	1	NA
teh	Tehuelche	Chon Proper	-68	-48	South America	4	4	1	1	35
tks	Teke (Southern)	Bantoid	14.5	-2.33	Africa	38787	3	3	2	28
tel	Telugu	South-Central Dravidian	79	16	Eurasia	69688278	4	2	1	43
tmn	Temein	Temein	29.42	11.92	Africa	10000	2	3	2	25
tne	Temne	Southern Atlantic	-13.08	8.67	Africa	1200000	2	3	2	25
ter	Tera	Biu-Mandara	11.83	11	Africa	100620	5	2	3	48
ttn	Tetun	Central Malayo-Polynesian	126	-9	SE Asia & Oceania	450000	1	2	1	19
tha	Thai	Kam-Tai	101	16	SE Asia & Oceania	20229987	3	3	3	30
tib	Tibetan (St. Spoken)	Bodic	91	30	SE Asia & Oceania	1261587	5	3	2	NA
tic	Ticuna	Ticuna	-70.5	-4	South America	41000	2	2	3	29
tgk	Tigak	Oceanic	150.8	-2.72	SE Asia & Oceania	6000	1	2	1	17
tgr	Tigré	Semitic	38.5	16.5	Africa	8e+05	4	2	1	33
try	Tiruray	South Mindanao	124.17	6.75	SE Asia & Oceania	50000	2	2	1	22
twn	Tiwa (Northern)	Kiowa-Tanoan	-105.5	36.5	North America	927	4	2	2	38
tiw	Tiwi	Tiwian	131	-11.5	Australia-New Guinea	1500	3	1	1	26
tlp	Tlapanec	Subtiaba-Tlapanec	-99	17.08	North America	54000	3	NA	3	30
tli	Tlingit	Tlingit	-135	59	North America	845	5	1	2	48
toa	Toaripi	Eleman	146.25	-8.33	Australia-New Guinea	23000	1	2	1	14
tol	Tol	Tol	-87	14.67	North America	350	3	2	1	28
ton	Tonkawa	Tonkawa	-96.75	30.25	North America	NA	2	2	1	25
tpa	Totonac (Papantla)	Totonacan	-97.33	20.33	North America	80000	2	1	1	22
tru	Trumai	Trumai	-53	-12	South America	78	3	2	1	24
tsi	Tsimshian (Coast)	Tsimshianic	-129	52.5	North America	800	5	1	1	41
tso	Tsou	Tsouic	120.75	23.5	SE Asia & Oceania	2127	2	2	1	21

ttu	Tsova-Tush	Nakh	45.5	42.5	Eurasia	3420	5	2	1	45
tug	Tuareg (Ahaggar)	Berber	6	23	Africa	62000	4	3	1	37
tuk	Tukang Besi	Sulawesi	123.5	-5.5	SE Asia & Oceania	250000	3	2	1	NA
tul	Tulu	Southern Dravidian	75.33	12.75	Eurasia	1949000	3	3	1	37
tun	Tunica	Tunica	-91	32.67	North America	NA	2	3	NA	24
tur	Turkish	Turkic	35	39	Eurasia	50625794	3	3	1	33
tuv	Tuvan	Turkic	95	52	Eurasia	209400	3	3	1	29
tza	Tzeltal (Aguacaten.)	Mayan	-92.5	16.42	North America	90000	3	2	1	28
umb	UMbundu	Bantoid	15	-12.5	Africa	4002880	3	2	2	NA
una	Una	Mek	140	-4.67	Australia-New Guinea	4000	3	3	2	NA
ung	Ungarinjin	Wororan	126	-16.33	Australia-New Guinea	82	3	2	1	24
urk	Urubú-Kaapor	Tupi-Guaraní	-46.5	-2.33	South America	500	2	2	1	NA
usa	Usan	Madang	145.17	-4.83	Australia-New Guinea	1400	1	2	1	20
uzn	Uzbek (Northern)	Turkic	66.5	40.67	Eurasia	18795591	3	2	1	30
vie	Vietnamese	Viet-Muong	106.5	10.5	SE Asia & Oceania	67439139	3	3	3	36
wah	Wahgi	Chimbu	144.72	-5.83	Australia-New Guinea	86000	2	2	2	23
wam	Wambaya	West Barkly	135.75	-18.67	Australia-New Guinea	12	2	1	1	NA
wnt	Wantoat	Finisterre-Huon	146.5	-6.17	Australia-New Guinea	8201	1	3	1	21
wps	Wapishana	Arawakan	-60	2.67	South America	7500	2	1	1	25
wap	Wappo	Wappo	-122.5	38.5	North America	NA	4	2	1	35
wra	Warao	Warao	-61.67	9.33	South America	18000	1	2	1	21
wry	Waray (in Australia)	Waray	131.25	-13.17	Australia-New Guinea	4	2	2	1	21
wrd	Wardaman	Yangmanic	131	-15.5	Australia-New Guinea	50	2	2	1	NA
war	Wari	Chapacura-Wanhan	-65	-11.33	South America	1833	2	2	1	NA
wrs	Waris	Border	141	-3.17	Australia-New Guinea	2500	1	3	NA	22
wma	West Makian	North Halmaheran	127.58	0.5	Australia-New Guinea	12000	2	2	1	23
wdo	W. Desert (Ooldea)	Pama-Nyungan	132	-30.5	Australia-New Guinea	NA	2	1	1	20
wch	Wichí	Matacoan	-62.58	-22.5	South America	15000	3	2	1	NA
wic	Wichita	Caddoan	-97.33	33.33	North America	3	3	1	1	29
wmu	Wik Munkan	Pama-Nyungan	141.75	-13.92	Australia-New Guinea	400	1	2	1	18
win	Wintu	Wintuan	-122.5	41	North America	5	4	2	1	35
wiy	Wiyot	Wiyot	-124.17	40.83	North America	NA	3	2	1	29
woi	Woisika	Timor-Alor-Pantar	124.83	-8.25	Australia-New Guinea	16522	2	3	1	28
wlf	Wolof	Northern Atlantic	-16	15.25	Africa	3612560	4	3	1	40
wuc	Wu (Changzhou)	Chinese	119.92	31.67	SE Asia & Oceania	77175000	4	3	3	34
xia	Xiamen	Chinese	118.17	24.5	SE Asia & Oceania	46227965	3	2	3	25
xoo	!Xóõ	Southern Khoisan	21.5	-24	Africa	4200	5	2	3	NA
ygr	Yagaria	Eastern Highlands	145.42	-6.33	Australia-New Guinea	21116	1	2	2	23
yag	Yagua	Peba-Yaguan	-72	-3.5	South America	5692	1	2	2	23
ykt	Yakut	Turkic	130	62	Eurasia	363000	3	3	1	34
yan	Yana	Yana	-122	40.5	North America	NA	3	2	1	30
yny	Yanyuwa	Pama-Nyungan	137.17	-16.42	Australia-New Guinea	70	4	1	1	32
yap	Yapese	Yapese	138.17	9.58	SE Asia & Oceania	6592	3	3	1	NA
yaq	Yaqui	Cahita	-110.25	27.5	North America	16406	2	2	2	22
yar	Yareba	Yareban	148.5	-9.5	Australia-New Guinea	750	1	2	1	18
yaw	Yawa	Yawa	136.25	-1.75	Australia-New Guinea	6000	1	2	1	19

yay	Yay	Kam-Tai	104.75	22.42	SE Asia & Oceania	2049203	3	2	3	34
yel	Yelî Dnye	Yele	154.17	-11.37	Australia-New Guinea	3750	5	3	1	NA
yes	Yessan-Mayo	Tama Sepik	142.58	-4.17	Australia-New Guinea	1988	2	1	1	20
yey	Yeyi	Bantoid	23.5	-20	Africa	25200	5	2	2	NA
yid	Yidiny	Pama-Nyungan	145.75	-17	Australia-New Guinea	12	1	1	1	16
yim	Yimas	Lower Sepik	143.55	-4.67	Australia-New Guinea	300	1	1	1	NA
yor	Yoruba	Defoid	4.33	8	Africa	19327000	2	3	3	29
yct	Yucatec	Mayan	-89	20	North America	7e+05	3	2	2	30
yuc	Yuchi	Yuchi	-86.75	35.75	North America	10	5	2	1	46
ycn	Yucuna	Arawakan	-71	-0.75	South America	1800	2	2	1	21
yko	Yukaghir (Kolyma)	Yukaghir	150.83	65.75	Eurasia	10	3	2	1	NA
ytu	Yukaghir (Tundra)	Yukaghir	155	69	Eurasia	30	3	2	1	26
yul	Yulu	Bongo-Bagirmi	25.25	8.5	Africa	7000	5	3	3	42
yus	Yupik (Siberian)	Eskimo-Aleut	-173	65	North America	1350	4	1	1	36
yur	Yurok	Yurok	-124	41.33	North America	12	4	2	1	NA
zan	Zande	Adamawa-Ubangian	26	4	Africa	1142000	3	3	2	35
zqc	Zoque (Copainalá)	Mixe-Zoque	-93.25	17	North America	10000	2	2	1	22
zul	Zulu	Bantoid	30	-30	Africa	9563422	4	2	2	37
zun	Zuni	Zuni	-108.83	35.08	North America	9651	3	2	1	25
mak	Makah	Southern Wakashan	-124.67	48.33	North America	NA	NA	2	1	NA
ngi	Ngiyambaa	Pama-Nyungan	145.5	-31.75	Australia-New Guinea	12	NA	1	1	18
tas	Tashlhiyt	Berber	-5	31	Africa	3e+06	NA	1	1	31
wao	Waorani	Waorani	-76.5	-1	South America	1650	NA	2	1	21

4. References

S1. Q.D. Atkinson, Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science*, **332**, 346-349 (2011).

S2. I. Maddieson, *Patterns of Sounds*. (Cambridge: Cambridge University Press, 1984).

S3. I. Maddieson, K. Precoda, The UCLA Phonological Segment Inventory Database. (1990).

S4. M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds., *The World Atlas of Language Structures* (Oxford, UK, Oxford Univ. Press, 2005).

S5. L. Hyman. 2008. Universals in phonology. *The Linguistic Review*, **25**, 83-137.

S6. N. Clements, The Role of Features in Phonological Inventories, in *Contemporary Views on Architecture and Representations in Phonology*, E. Raimy, C. E. Cairns, Eds. (MIT Press, 2009), 19-68.

S7. I. Maddieson. Consonant inventories, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 10-13.

S8. I. Maddieson. Vowel quality inventories, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 14-17.

S9. I. Maddieson. Tone, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 58-61.

S10. H. Reetz. Web interface to UPSID. http://web.phonetik.uni-frankfurt.de/upsid_info.html.

S11. T. Güldemann, Phonological regularities of consonant systems across Khoisan lineages. *University of Leipzig Papers on Africa: Language and Literatures* **16** (2001).

S12. H. Nakagawa. Aspects of the phonetic and phonological structure of the G/ui language. (Johannisburg: PhD Thesis University of the Witwatersrand, 2008).

S13. M. S. Dryer, M. Haspelmath, Eds., *The World Atlas of Language Structures Online.* (Munich: Max Planck Digital Library, 2011). http://wals.info/.

S14. D. Dediu, D.R. Ladd, Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 10944-10949 (2007).

S15. J. Hajek. Vowel nasalization, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 46-49.

S16. M. P. Lewis, Ed., *Ethnologue: Languages of the World, Sixteenth edition* (Dallas: SIL International, 2009).

S17. D. Nettle. Is the Rate of Linguistic Change Constant? *Lingua* **108,** 119-136 (1999).

S18. P. Trudgill, Linguistic and Social Typology: The Austronesian Migrations and Phoneme Inventories. *Linguistic Typology*, **8**, 305-320 (2004).

S19. J. Hay, L. Bauer, Phoneme Inventory Size and Population Size. *Language*, **83**, 388-400 (2007).

S20. W. Lehfeldt, Die Verteilung der Phonemanzahl in den natürlichen Sprachen. *Phonetica*, **31**, 247-287 (1975).

S21. J. Justeson, L. Stephens, On the Relationship Between the Numbers of Vowels and Consonants in Phonological Systems. *Linguistics*, **22**, 531-545 (1984).

S22. M. Cysouw, On the Probability Distribution of Typological Frequencies, in *The Mathematics of Language*, C. Ebert, G. Jäger, J. Michaelis, Eds. (Berlin: Springer, 2010), pp. 29-35.

S23. S. Mithen. *After the Ice* (London, Orion Books, 2003).

S24. S. Wichmann, E W. Holman, Population size and rates of language change. *Human Biology* **81**, 259--274 (2009).

S25. J. Diamond, P. Bellwood, Farmers and Their Languages: The First Expansions. *Science* **300**, 597-603 (2003).

S26. M. S. Dryer, Large Linguistic Areas and Language Sampling. Studies in Language **13**, 257-92 (1989).

S27. P.V. Kirch, Peopling of the Pacific: a holistic anthropological perspective. *Annu. Rev. Anthropol*. **39**, 131-148 (2010).

S28. A. Manica, W. Amos, F. Balloux, T. Hanihara, The effect of ancient population bottlenecks on human phenotypic variation. *Nature* **448**, 346-348 (2007).

S29. L. Betti, F. Balloux, W. Amos, T. Hanihara, A. Manica, Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc Biol Sci* **276**, 809-814 (2009).

S30. G.E. Schwarz, Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464 (1978).

S31. J. H. McWhorter. The world's simplest grammars are creole grammars. *Linguistic Typology* **5**, 125–166 (2001).

S32. G. Sampson, D. Gil, P. Trudgill, Eds., *Language complexity as an evolving variable*. (Oxford: Oxford University Press, 2009)

S33. G. Lupyan, R. Dale. Language Structure Is Partly Determined by Social Structure. *PLoS ONE* **5**, e8559, (2010).

S34. G. D. S. Anderson. The Velar Nasal, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 42-45.

S35. I. Maddieson. Syllable Structure, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 54-57.

S36. B. Bickel, J. Nichols, Inflectional Synthesis of the Verb, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 94-97.

S37. C. Rubino. Reduplication, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 114-117.

S38. G. G. Corbett. Number of Genders, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 126-129.

S39. M. Haspelmath, Occurrence of Nominal Plurality, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 142-145.

S40. H. Diessel, Distance Contrasts in Demonstratives, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 170-173.

S41. E. König, P. Siemund, S. Töpper, Intensifiers and Reflexive Pronouns, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 194-197.

S42. O. Iggesen, Number of Cases, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 202-205.

S43. D. Gil, Numeral Classifiers, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 226-229.

S44. J. Nichols, B, Bickel, Possessive Classification, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 242-245.

S45. Ö. Dahl, V. Velupillai, Perfective/Imperfective Aspect, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 267-268.

S46. Ö. Dahl, V. Velupillai, The Future Tense, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 270-271.

S47. F. de Haan, Semantic Distinctions of Evidentiality, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 314-317.

S48. A. Siewierska, Passive Constructions, in *The World Atlas of Language Structures,* M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Oxford, UK, Oxford Univ. Press, 2005), pp. 434-437.

S49. D. Bates, M. Maechler, lme4: Linear mixed-effects models using S4 classes. *R package,* version 0.999375-37. (2010). http://CRAN.R-project.org/package=lme4.

S50. K. P. Burnham, D. R. Anderson, *Model Selection and Inferences* (Springer, New York, 1998).

S51. D.R. Anderson, K.P. Burnham, G.C. White, Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture–recapture studies. *J. Appl. Stat.* **25**, 263–282 (1998).

S52. R Development Core Team, R: A language and environment for statistical computing. (Vienna: R Foundation for Statistical Computing, 2010). http://www.R-project.org.

S53. R. H. Baayen, languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". *R package*, version 1.0. (2010). http://CRAN.R-project.org/package=languageR

S54. H. Akima, A. Gebhardt, T. Petzoldt, M. Maechler, akima: Interpolation of irregularly spaced data. *R package,* version 0.5-4. (2009). http://CRAN.R-project.org/package=akima.

S55. R. Turner, deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation. *R package*, version 0.0-13. (2010). http://CRAN.R-project.org/package=deldir

S56. C. T. Butts, sna: Tools for Social Network Analysis. *R package*, version 2.2-0. (2010). http://CRAN.R-project.org/package=sna

S57. R. Furrer, D. Nychka and S. Sain, fields: Tools for spatial data. *R package*, version 6.3. (2010). http://CRAN.R-project.org/package=fields

S58. C. Renfrew, A.M. McMahon, L. Trask, Eds., *Time Depth in Historical Linguistics* (Cambridge, UK, The McDonald Institute for Archaeological Research, 2000).

S59. M. Dunn, A. Terrill, G. Reesink, R.A. Foley, S.C. Levinson, Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072-2075 (2005).

S60. D. Dediu, A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proc. R. Soc. London Ser. B* **278**, 474-479 (2011).

S61. Wichmann, S., E. Holman. *Assessing Temporal Stability for Linguistic Typological Features.* (München: LINCOM Europa, 2009)

S62. J. Relethford, *Reflections of our past: How human history is revealed in our genes* (MA: Westview Press, 2003).

S63. V. Eswaran, H. Harpending, A.R. Rogers, Genomics refutes an exclusively African origin of humans. *J. Hum. Evol.* **49**, 1-18 (2005).

S64. A. Templeton, Out of Africa again and again. *Nature*, **416**, 45-51 (2002).

S65. S.G.Thomason, T. Kaufman, Language contact, creolization, and genetic linguistics. (Berkeley, Univ California Press, 1988).

S66. T.F. Jaeger, P. Graff, W. Croft, D. Pontillo. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* **15**, 281-319 (2011)

S67. I. Maddieson, T. Bhattacharya, D.E. Smith and W. Croft. Geographical distribution of phonological complexity. *Linguistic Typology* **15**, 267-279 (2011).

S68. C. Bowern. Out of Africa? The logic of phoneme inventories and founder effects. *Linguistic Typology* **15**, 207-216 (2011).