Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"

Rory Van Tuyl[1]* and Asya Pereltsvaig[2]

Atkinson (Reports, 15 April 2011, p. 346) concluded that language originated in western Africa and that, due to a serial founder effect, languages repeatedly lost phonemes the farther they moved from the African point of origin. Independent examination of the published data tends to refute both these claims.

Recently, Atkinson (*1*) plotted total normalized phoneme diversity (TNPD), a synthetic measure of language complexity, against distance from a putative point of origin (PO) in western Africa. By forcing a linear least-squares fit to the data and adjusting the PO location, he observed a maximum negative correlation when the PO was set to latitude 1°18′S, longitude 9°46′E on the equatorial west coast of Africa (*2*). He interpreted this maximum-correlation linear regression fit as evidence for a serial founder effect (SFE), a hypothesis that assumes small bands of isolated pioneers repeatedly tend to develop languages with fewer phonemes than that of their progenitors, and he interpreted the location of the PO as evidence that language most likely originated in western Africa.

A plot of data (*3*) for the human migration path from Africa to South America (Fig. 1) shows a better fit when data are linearly regressed continent by continent, revealing none of the downward slope associated with an SFE for any continent but Africa (*4*). An SFE must surely have existed in settling the vast expanse of Eurasia and the Americas, but there seems to be no evidence of it in these data. Why then should an SFE be considered the cause of TNPD regression slope within Africa?

Because distance was measured through fixed waypoints between continents, two-dimensional interaction between the PO and data points outside Africa was lost, so adjusting the location of the PO has no effect on the correlation for data outside Africa [Fig. 1 and supporting online material (SOM)]. Therefore, one would expect the PO to have been determined based solely on African data, and it was. When correlation is maximized using African data only, the PO changes negligibly, demonstrating that non-African data are irrelevant to determination of the extrapolated point of maximum phoneme diversity within Africa.

Linear regression centered at the western African PO is weak and may be adventitious. Such trends can occur purely by chance in meaningless locations. For example, the point of maximum correlation for the north Asian data is in northern Burma (an unlikely origin for human language), and for Nilo-Saharan languages it is far south of their range, in the middle of the Congo rainforest. Thus, an extrapolated PO based on maximum correlation can signal something other than a logical point of language origin, and the zone of putative western African language origin shown in Atkinson's figure 2A may be an artifact of analysis. Our Fig. 2 shows that correlation contours surrounding the PO may largely be due to

[1]Independent Scholar, Los Altos, CA 94022, USA. [2]Department of Linguistics, Stanford University, Stanford, CA 94305, USA.

*To whom correspondence should be addressed. E-mail: roryvantuyl@gmail.com

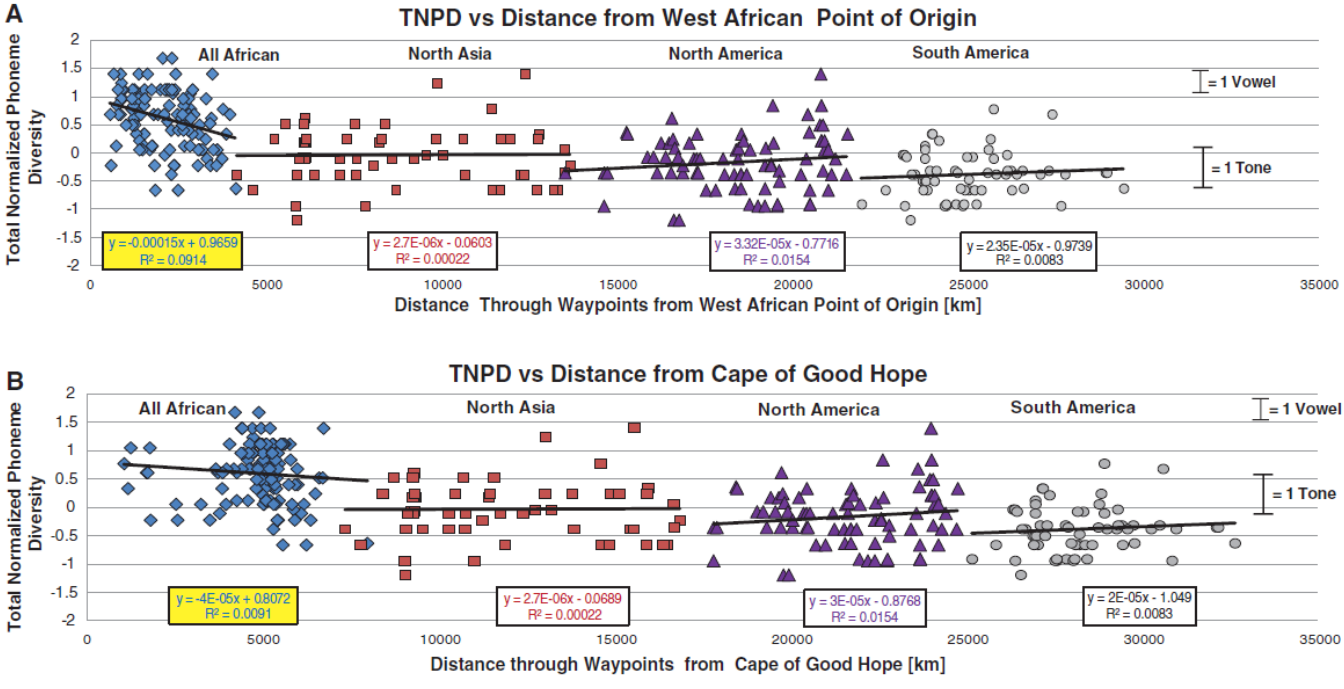

Fig. 1. TNPD for sub-Saharan Africa, north Asia (latitude > 30°, longitude > 30°), North America, and South America, versus (**A**) distance from the African point of origin and (**B**) distance from Cape of Good Hope. Continents outside Africa show no continuous phoneme loss versus distance. Africa shows a maximum decline (A) of about 1 SD of the data (0.48) with poor correlation (R^2 = 0.091) when a western African PO is chosen (A) and little decline for a southern African PO (B). When the PO within Africa is changed, the slopes and correlations for the African data change but non-African continents remain constant, showing that they have no active contribution to determining a maximum-correlation PO within Africa (see SOM).

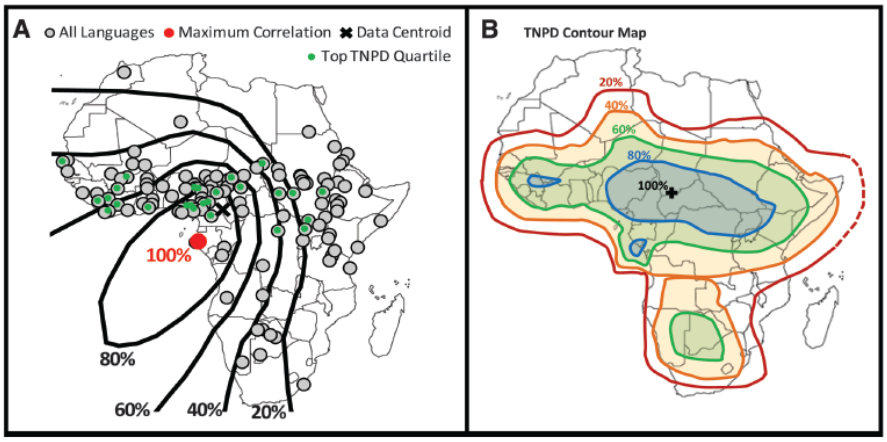

Fig. 2. (**A**) Map of African languages with contours of constant normalized correlation coefficient centered about a peak value (100%). These contours are similar to those for "likely area of language origin" in Atkinson's figure 2A. Language data are mostly in an equatorial swath, with fewer data to the south and very few to the north. (**B**) Map showing contours of normalized TNPD values (1000-km resolution). We see that the correlation contours in (A) do not accurately portray the distribution of TNPD (which stretches across the continent at about 7.5° north latitude) or the actual TNPD data peaks. Instead, the correlation method indicates elevated regions of correlation coefficient well into the Atlantic Ocean. The TNPD contour map shows peak TNPD near the data centroid, evidently a direct result of the data's geographic distribution.

the way the correlation technique interacts with the geographic distribution of data points, regardless of data's values (see SOM).

In addition to the above data analysis problems, there are reasons to doubt the data itself. TNPD as a measure of language complexity is highly problematic with regard to how phonemes are counted and how different phoneme types are combined. There are significant variations in phoneme counts reported from different sources, especially with respect to tones (*5*). TNPD calculations involve all consonant phonemes, some vowel phonemes (phonemic distinctions based on nasality and length are ignored), and tone features (tones are a suprasegmental feature on a par with length or stress and not a phoneme per se). Furthermore, assigning numerical values to the *World Atlas of Language Structures* (WALS) categories is questionable: In the TNPD scoring system, each vowel is on average worth 2.6 consonants, each tone worth 5.7 consonants. These ratios seem arbitrary, as is averaging vowels, consonants, and tones to calculate TNPD. As a result, consonant inventory plays virtually no role in the regression versus distance, and tones dominate the correlation (*6*). Furthermore, the TNPD measure also conceals that languages typically acquire or lose not single phonemes but whole (natural) classes of phonemes, such as clicks or ejectives, long or nasal vowels, or tonal distinctions.

Difficulties arise when a single numerical score is assigned to each WALS category, because with quantized categories the addition or subtraction of but a single phoneme can result in the same TNPD change as with multiple phonemes. Small changes in phoneme counts can have a substantial effect on analytical outcome. We conducted a study of 10 sub-Saharan languages that showed a discrepancy in 11 out of 30 counts (three for each language) between the WALS data and other sources, resulting in 8 of 10 languages changing TNPD quartile and the regression slope disappearing when the non-WALS data were used (*7*). [One African language, despite having 122 consonants (*8*), ranked in the second TNPD quartile.]

An SFE presupposes that languages change incrementally, in isolation from their neighbors. However, it is hard to see how the number of phonemes in a given language can be unambiguously attributed to a founder effect because languages are known to change in response to influences from neighbors (e.g., some Bantu languages acquired click sounds under the influence of Khoisan languages, and some Indo-Aryan languages acquired retroflex consonants from their Dravidian neighbors). Furthermore, DNA research argues for modern human origin in either eastern Africa (*9*) or southern Africa (*10*), and we can assume language originated with these people. Whatever the locations and phoneme inventories were for African languages in antiquity, the situation is surely different today, some 50,000 years after the modern human exodus. Migrations, conquests, and borrowings—many of which occurred long after the era of the founder effect—can explain the present state of African languages more credibly than simple diffusion of small founder groups.

References and Notes

1. Q. D. Atkinson, *Science* **332**, 346 (2011).
2. Triangulation of published data.
3. www.sciencemag.org/cgi/content/full/332/6027/346/DC1.
4. Overall fit to data is about 7% better to the segmented trend line shown (s = 0.473) than to an overall linear trend line (s = 0.505). The differences between mean values for adjacent continents are statistically significant between African and all non-African languages (t test; P = 0.000) and between North and South American (t test: P = 0.004), but not between north Asian and
5. I. Maddieson, in *The World Atlas of Language Structures Online*, M. S. Dryer, M. Haspelmouth, Eds. (Max Planck Digital Library, Munich, 2011); chaps. 1, 2, and 13 (accessed 5/21/2011); http://wals.info.
6. Tones make up 59%, vowels 40%, and simple consonants 1% of the African regression slope.
7. Alternate data were obtained from Wikipedia Articles (accessed 5/23/2011) on the following languages: Swahili, Zulu, Jul'hoan, Sandawe, Khoekhoe, Maasai, Dinka, Luo, Igbo, and Yoruba. We do not assert these data to be correct; we merely note that they are different from the WALS data and reveal variation in phoneme counts between sources. Alternate data for the entire language set were not available and may not exist.
8. The language is !Xóõ. It has more than five times the average number of consonants, but fewer than five vowels [WALS (*5*), chapter 1].
9. D. M. Behar *et al.*; Genographic Consortium, *Am. J. Hum. Genet.* **82**, 1130 (2008).
10. B. M. Henn *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5154 (2011).

Supporting Online Material

www.sciencemag.org/cgi/content/full/335/6069/657-d/DC1
Materials and Methods
Figs. S1 and S2
Table S1
References

1 June 2011; accepted 3 January 2012
10.1126/science.1209176