Does phoneme inventory size correlate with population size?

MARK DONOHUE and JOHANNA NICHOLS

1. Introduction and method

Atkinson 2011 finds a significant positive correlation between population size and phoneme inventory size (confirming Hay & Bauer 2007) and explains it by migration: phoneme sizes are largest in Africa, and as societies spread out of Africa and around the world they went through population and cultural bottlenecks and underwent phonological simplification as a consequence. We believe the correlation is artefactual if it exists at all, and it probably does not exist.

To test it we surveyed 1,350 languages with excellent genealogical and geographical coverage and distribution. We used the Autotyp areal breakdown (Bickel & Nichols 2002). Population size figures were taken, where possible, from grammars, ethnographies, and/or recent census data, and attempt to give population figures for the entire ethnic group (and not just speakers of the language, since most of the world's languages are losing speakers to large national and international languages); where we did not have this information we used figures from Lewis (ed.) 2009. Language shift has increased rapidly in recent decades, so that the size of the ethnic group is a better measure of the size of the speech community in which the oldest and most fluent speakers grew up, and it is these speakers' competence that grammars usually describe. The log of population size was coded for each language. Since ethnic groups and speech communities of under a few hundred individuals are ordinarily unstable (not impossibly, but the languages of such communities are frequently undergoing shift and death), all population sizes reported in units or tens were coded as 499 (treating these speech communities as though they still had the sizes reported for them in the early to mid-twentieth century).

For each sample language we surveyed the total number of consonant phonemes (excluding phones found only in unassimilated foreign loans), the total number of phonemic vowel qualities, and the number of tone oppositions.

Linguistic Typology 15 (2011), 161–170
DOI 10.1515/LITY.2011.011
1430–0532/2011/015-0161

(The vowel count excludes any diphthongs amenable to a bisegmental analysis. Thus the diphthongized vowels of English *bait* and *boat* are included as vowel phonemes, but the diphthongs of *bite*, *butte*, *bout*, and *Boyd* are excluded as bisegmental by some criteria.) Phonemic length and nasalization were not included; this lessens the accuracy of the survey but it proved impossible to add in this additional information before the deadline for submitting comments. Contrastive registers and phonation types are also not included, for the same practical reasons.

Thus in its data choice this survey resembles Atkinson's, but neither in the sample design nor in the data coding do we replicate Atkinson's method. Rather, we attempt a preliminary and approximate answer to the question of whether any correlation exists between population size and phoneme inventory size. We deliberately chose to employ simple, easily replicated statistical methods, since the nature of the population data available is such that more sophisticated techniques are not warranted.

2. No worldwide correlation

Within the entire sample there is no significant correlation between magnitude of population and size of phoneme inventory ($r = 0.27$). See Figure 1, which displays the results for a 1,350 language sample; trendline is nearly level. (We obtained very similar results in a pilot survey of a genealogically and areally balanced sample of 286 languages.)

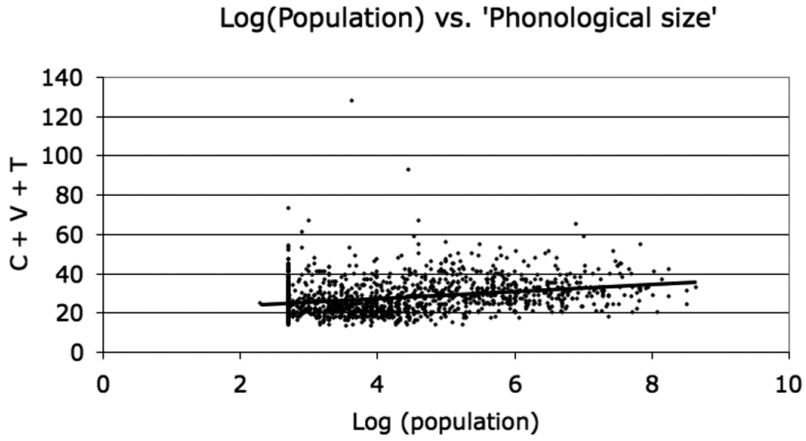

Figure 1. *Log of population and number of phonemes worldwide. N* = 1,350*. The outliers are !Xóõ and Ju|'hoan, click languages of southern Africa.*

Table 1. *Areal means for population magnitude and two measures of phoneme inventory size: consonants alone (Cons.), and the sum of consonants, vowel qualities, and tones (All). N = number of languages. Based on the smaller but geographically more balanced Autotyp sample (Bickel & Nichols 2002). (Three languages were omitted from Western-Southwestern Eurasia for this count, because their population magnitudes were identical to the mean.) Areas are arranged in approximate order of increasing distance from Africa.*

	Population	Cons.	All	N
Africa	5.76	29.64	38.21	39
Western-Southwestern Eurasia	6.50	30.25	36.50	17
North-Central Asia	5.63	20.74	28.05	19
South/Southeast Asia	6.06	26.59	34.34	32
New Guinea-Oceania	4.31	15.08	21.74	53
Australia	3.18	17.94	21.73	33
Western North America	3.19	27.85	32.96	26
Eastern North America	3.61	20.47	26.32	19
Central America	4.63	19.08	24.83	12
South America	4.36	17.33	23.33	33
Total				283

But when the sample languages are aggregated into continent-sized areas (Table 1), a roughly east-to-west correlation emerges. Mean populations for the sample languages are largest in Africa and Eurasia, smaller in New Guinea-Oceania, Central America, and South America, and smallest in Australia and North America, and phoneme inventory sizes show a similar cline. This is evidently the correlation detected by Atkinson and by Hay & Bauer. Below it is argued that this correlation is artefactual, not representing an intrinsic fact about language but reflecting the different political and economic histories of different continents over the last two millennia (discussed again below).

3. No correlations within areas

Within the same large areas, no positive correlation of population size with phoneme inventory size is evident (Table 2). There is a significant negative correlation in western and southwestern Eurasia, mostly due to the fact that this area includes western Europe with its several large, phonologically simple national languages (e.g., English, Spanish) and the Caucasus with its many small, phonologically complex languages. There are slight negative skewings in Africa and in South-Southeast Asia, and a slight positive skewing in New Guinea-Oceania, none approaching significance. (We obtained very similar results on a survey of just consonant inventories.)

Table 2. *Areal correlations between log(population) and phoneme inventory size. Entries are numbers of languages. Significance levels from Fisher's exact test (two-tailed).*

	Population	Languages with above-mean and below-mean phoneme inventories		*p*
		Below	Above	
Africa	Below	9	6	
	Above	18	6	0.4775
Western-Southwestern Eurasia	Below	1	8	
	Above	8	0	0.0004
North-Central Asia	Below	6	5	
	Above	4	4	1.0000
South/Southeast Asia	Below	8	10	
	Above	10	4	0.1649
New Guinea-Oceania	Below	15	12	
	Above	11	15	0.4142
Australia	Below	13	14	
	Above	3	3	1.0000
Western North America	Below	10	11	
	Above	2	3	1.0000
Eastern North America	Below	9	2	
	Above	4	4	0.3189
Central America	Below	3	2	
	Above	4	3	1.0000
South America	Below	10	8	
	Above	8	7	1.0000

On Atkinson's model, the bottlenecks and isolation accounting for decreased phoneme inventory size occurred during the upper Paleolithic expansion of modern humans to Asia and from there to the Pacific, the high latitudes, and the Americas. As Sproat (2011a, b) points out, all populations at that time must have been small, so if an effect of population size on phoneme inventory size is real it should be evident among populations numbering in the hundreds and low thousands. In our sample it should be evident in New Guinea-Oceania, Australia, and the Americas (which have the lowest mean populations in Table 1), but there is no evidence for it there. A plot like that in Figure 1, but only for populations up to 5,000, has an almost exactly level slope ($R^2 = 0.0077$; this plot is not reproduced here).

4. No correlations within families

Within those large families for which we have adequate sampling, no consistent correlation of population size with phoneme inventory size can be detected (Figure 2). There is no appreciable correlation in some families (e.g., Tibeto-Burman and Austronesian, $r = 0.05$ and 0.03, for $n = 89$ and 220, respectively); weak but insignificant correlations in, for instance, Trans New Guinea ($r = -0.12$, $n = 92$) and Indo-European ($r = 0.10$, $n = 82$); significant NEGATIVE correlation in families such as Uto-Aztecan ($r = -0.37$, $n = 21$) and Algic ($r = -0.37$, $n = 10$); significant POSITIVE correlation in some families, such as Dravidian ($r = 0.39$, $n = 15$) and Mayan ($r = 0.34$, $n = 12$). There is no region of the world where an unambiguous trend can be found. Scatterplots for the families discussed in this paragraph are shown in Figure 2; Table 3 presents additional summary data from additional families.

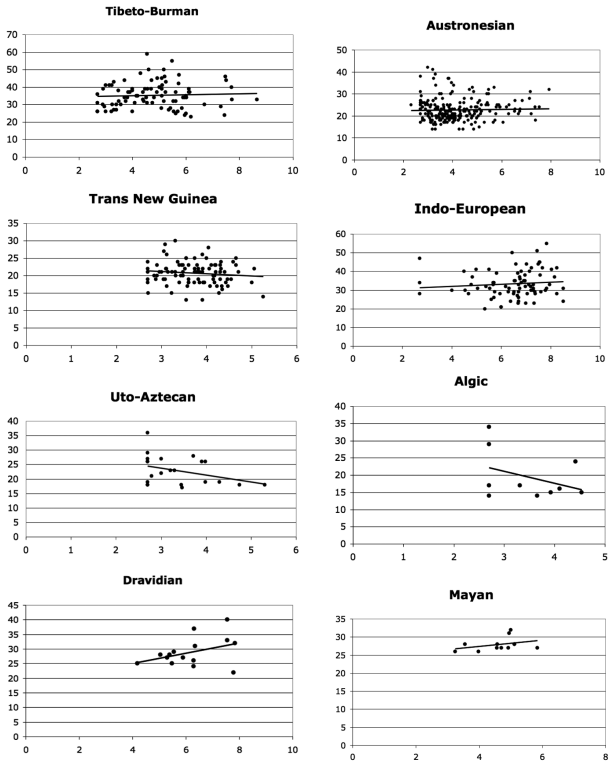

Figure 2. *Log of population and number of phonemes in eight families worldwide*

Table 3. *Genealogical correlations between log(population) and phoneme inventory size. The correlation is given as* r*; numbers of languages in the sample for each family considered is shown as* n*. Only families or macrofamilies for which our database has at least 5 members were considered.*

	Family	*r*	*n*
Africa	Afro-Asiatic	−0.21	43
	Khoe	−0.20	6
	Mande	0.04	11
	Niger-Kongo	0.16	103
	Ubangi	−0.44	7
Western-Southwestern Eurasia	Indo-European	0.10	82
	Nakh-Daghestanian	0.29	16
	Turkic	0.30	18
North-Central Asia	Dravidian	0.39	15
	Mongolic	0.18	8
	Tibeto-Burman	0.05	89
	Tungusic	−0.76	6
	Uralic	−0.12	14
South/Southeast Asia	Austronesian	0.03	220
	Austro-Asiatic	−0.30	28
	Tai-Kadai	−0.12	10
	Timor-Alor-Pantar	0.08	10
New Guinea-Oceania	Lower Sepik-Ramu	0.53	6
	Sepik	0.13	10
	Skou	−0.26	7
	Trans New Guinea	−0.12	92
	West Papuan	0.04	12
Australia	Gunwinyguan	0.34	8
	Pama-Nyungan	0.18	73
Western North America	Eskimo-Aleut	−0.59	5
	Na-Dene	0.30	11
	Salishan	−0.23	8
	Yuman	−0.41	7
Eastern North America	Algic	−0.37	10
	Iroquoian	0.23	5
	Siouan	0.38	7
Central America	Mayan	0.34	12
	Otomanguean	0.64	5
	Uto-Aztecan	−0.37	21
South America	Arawak	−0.25	14
	Carib	−0.48	7
	Chibchan	0.13	10
	Macro-Je	−0.42	8
	Panoan	0.31	9
	Tucanoan	0.13	8
	Tupí	0.75	9

5. Discussion

5.1. *Explaining the parallel east-west clines*

5.1.1. *Population sizes.* The roughly west-to-east cline of decreasing population size is due to the long history of statehood and empire in Africa and Eurasia (such political systems spread state languages at the expense of smaller ones), economic growth, and efficient food production (themselves accidents of geography: Diamond 1997). States and empires cause language spreading and lower linguistic diversity (Austerlitz 1980, Nichols 1992, Nettle 1999) and thereby create large speech communities. The smallest mean populations in our sample are found in the chiefly or entirely foraging areas (Australia, the Americas); intermediate levels occur where there was a mix of foraging and small farming or horticulture (Oceania, New Guinea, parts of eastern North America, eastern South America).

Furthermore, European colonization brought smallpox and economic destruction to the Americas and Australia, drastically reducing populations, and the population figures we have are post-colonial.

5.1.2. *Complexity.* Big trade languages, state languages, and other interethnic languages tend to be simpler than small ethnic languages (Trudgill 2009, Szmrecsanyi & Kortmann 2009, Dahl 2004), and there have been many more such large-population languages in Africa and Eurasia than in the pre-contact Americas and Pacific.

5.2. *Evidence against a causal correlation between population size and phoneme inventory size*

If sheer distance of migration, and the isolation and bottlenecks it implies, regularly lowered phonological complexity, then one would expect concomitant simplification of morphology and the rest of grammar with greater distance from Africa. In fact, however, we find the reverse: languages in the Americas and the Pacific are on average more complex overall, and morphologically, than those in Africa (or Africa plus Eurasia) (Nichols 2009). Consequently, there is a highly significant negative correlation between overall complexity and population size, which shows up on a merged worldwide sample but is not visible within areas. The fact that this (artefactual) correlation is negative while the phonological one found by Atkinson is positive provides further evidence that migration distance is not responsible for either correlation.

A more fine-grained areal breakdown than we have used would reveal subcontinental clusters of complex phonologies in various worldwide hotspots, some of them close to Africa (notably the Caucasus), some more distant (the eastern Himalayas/highland Southeast Asia, (north-)western North America,

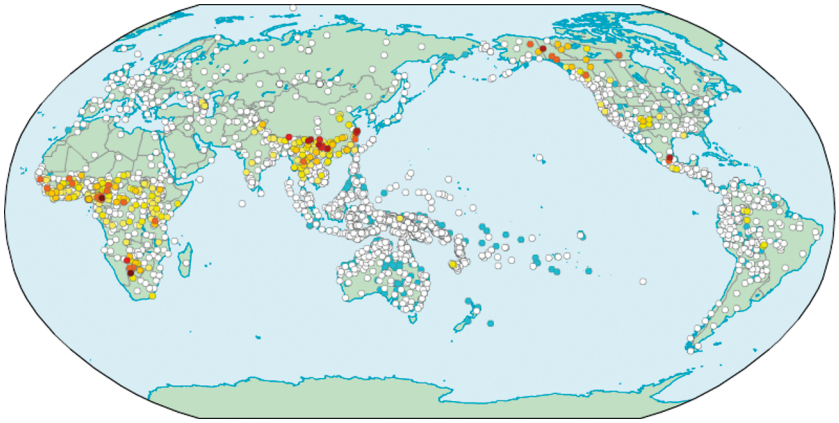

Figure 3. *Local hotspots for phonological complexity. Complexity calculated by multiplying total number of consonants × total syllable nuclei (contrastive vowel qualities + nasal vowel contrasts + length contrasts + phonation contrasts) × tonal contrasts. White dots are within one standard deviation of the mean; yellow, orange and (dark) red dots indicate increasing levels of elaboration; blue dots are very simple systems.*

Central America, northern Pre-Andine South America). In Figure 3 (based not on our genealogical-areal sample but on a larger survey of 1,357 languages and a larger set of phonological properties) we can see that some phonological systems in these areas are just as complex as those in Africa. A more sophisticated approach to phonological complexity would reveal still more local centers of complexity, such as the consonantal elaborations in the South American Andes, and the phonotactic freedoms encountered in western Eurasia (Europe and the Caucasus).

Atkinson's model does not predict these local hotspots, and in fact would seem to require a more or less monotonic decrease in complexity with distance from Africa. In fact, though, local areality is much more varied. Indeed, Africa is not itself uniformly complex, with at least two clearly identifiable "regions" with high complexity, the sub-Saharan belt (particularly west Africa, most particularly eastern Nigeria) and the Kalahari desert area.

Further evidence of the variability, and also the randomness, of areal phonological profiles, is shown by comparing Australia and Africa. These are two closed spread zones each of which has near-continental phonological areality as a result. In each of them an eccentric phonological type is widespread, and the types are diametrically opposed: great complexity of consonant inventories, especially of airstream mechanisms, is common in Africa while great simplicity predominates in Australia. Tones are ubiquitous in sub-Saharan Africa, categorically lacking in Australia.

Finally, there is a body of case-by-case evidence that isolation does not cause simplification. A well-known example is Kayardild, the aboriginal language of Bentinck Island, Australia, a small community that probably never numbered over a few hundred speakers. Despite being isolated for centuries or longer and having undergone population bottlenecks beginning pre-contact, Kayardild has a very standard phoneme inventory for Australia (Evans 1995). Another case is Rapanui, isolated for centuries on Easter Island; its sound system is much like that of its well-connected sisters in central and western Polynesia. What little can be gleaned about the sound system of Tasmanian (Crowley & Dixon 1981) suggests that it had a fairly typical Australian sound system despite millennia of isolation. Repeated anecdotal evidence comes from the indigenous languages of the Americas and Australia. These communities may have lost to 90 % of their population on European contact due to disease, warfare, and economic deprivation. Yet evidence from comparative reconstruction, colonial records, etc. does not indicate phonological simplification, or indeed any grammatical reflex of the population loss. There are also cases indicating that geographical isolation and its sociolinguistic analog, lack of inter-ethnic use, far from causing simplification, favor growth of complexity (Ross 1996, Nichols forthcoming).

6. Conclusion

A positive correlation between population size and size of phoneme inventory is critical to Atkinson's argument, but such a correlation is not expected given current knowledge of sociolinguistics, typology, and historical linguistics, and it cannot be demonstrated crosslinguistically.

Revised: 13 July 2011
Received: 20 June 2011

Australian National University
University of California, Berkeley

Correspondence addresses: (Donohue) Linguistics, ANU College of Asia & the Pacific, Australian National University, Canberra ACT 0200, Australia; e-mail: mark@donohue.cc; (Nichols) Department of Slavic Languages and Literatures, University of California at Berkeley, 6303 Dwinelle Hall #2979, Berkeley, CA 94720-2979, U.S.A.; e-mail: johanna@berkeley.edu

References

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

Austerlitz, Robert. 1980. Language-family density in North America and Eurasia. *Ural-Altaische Jahrbücher* 51. 1–10.

Bickel, Balthasar & Johanna Nichols. 2002. The Autotyp research program. http://www.uni-leipzig.de/~autotyp/

Crowley, Terry & R. M. W. Dixon. 1981. Tasmanian. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, Vol. 2, 394–421. Canberra: Australian National University Press.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: Benjamins.
Diamond, Jared. 1997. *Guns, germs, and steel: The fates of human societies*. New York: Norton.
Evans, Nicholas D. 1995. *A grammar of Kayardild*. Berlin: Mouton de Gruyter.
Hay, Jennifer & Laurie Bauer. 2007. Phoneme inventory size and population size. *Language* 83. 388–400.
Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world.* 16th edn. Dallas: SIL International.
Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.
Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.
Nichols, Johanna. 2009. Linguistic complexity: A comprehensive definition and survey. In Sampson et al. (eds.) 2009, 110–125.
Nichols, Johanna (forthcoming). The vertical archipelago: Adding the third dimension to linguistic geography. In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: Geographical, interactional, and cognitive perspectives.* Berlin: de Gruyter.
Ross, Malcolm D. 1996. Contact-induced change and the comparative method: Cases from Papua New Guinea. In Mark Durie & Malcolm D. Ross (eds.), *The comparative method reviewed,* 180–217. New York: Oxford University Press.
Sampson, Geoffrey, David Gil & Peter Trudgill (eds.). 2009. *Language complexity as an evolving variable.* Oxford: Oxford University Press.
Sproat, Richard. 2011a. *Science* does it again. Manuscript, Oregon Health & Science University. http://www.cslu.ogi.edu/~sproatr/newindex/atkinson.html (accessed 14 April 2011)
Sproat, Richard. 2011b. Phonemic diversity and the out-of-Africa theory. *Linguistic Typology* 15. 199–206.
Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. Between simplification and complexification: Non-standard varieties of English around the world. In Sampson et al. (eds.), 65–79.
Trudgill, Peter. 2009. Sociolinguistic typology and complexification. In Sampson et al. (eds.), 98–109.