On the geographical distribution of phoneme inventories

Michael Cysouw & Steve Moran, LMU München
First Draft. Please contact <cysouw@lmu.de> for any suggestions.

1. Introduction

[TODO: summarize Atkinson's article]

We see two main problems with the analysis as performed by Atkinson. First, his assessment of phoneme inventory size strongly emphasises tone and vowel distinctions in expense of consonant distinctions (Section 3). Second, there is an observed correlation between the size of the phoneme inventory and the size of the speaker community of a language (a correlation reiterated by Atkinson himself). When speaker community size indeed influences phoneme inventory size, then there is a problem, because there is a strong asymmetry in the geographical distribution of languages with large speaker communities. Languages with large speaker communities are predominantly found in Africa and Eurasia (Section 4). When both these considerations are included into the analysis, then does not seem to be much evidence anymore for an 'out-of-africa' language dispersal in the phoneme data (Section 5).

Note that in criticizing Atkinson's results we are not criticizing the insight of Africa as being the homeland of modern humans. We are only doubting Atkinson's proposal that this homeland can still be discerned in the modern distribution of phoneme systems across the world's languages. We would actually highly welcome any such conclusion on the basis of purely linguistic data, but we consider the current case to be—unfortunately—not convincing.

[TODO: geographical distance measurement as another possible problem]

[TODO: impact of internal structure of phoneme systems]

2. The Phoible database

[TODO: describe data collection and structure of database]

3. Estimating phoneme inventory size

The data used by Atkinson are taken from the *World Atlas of Language Structures* (WALS, (Haspelmath *et al* 2005)). Specifically, he combines the features "consonant inventories" (Maddieson 2005a), "vowel quality inventory" (Maddieson 2005c) and

“tone” (Maddieson 2005b) to obtain an estimate of the size of the phoneme inventory. Unfortunately, there are various idiosyncracies in the coding of the WALS data that influence Atkinson's results. First, WALS distinguishes rough classes of phoneme inventory size instead of actual numbers of phonemes. Second, Atkinson uses consonants, vowels and tone as equal 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 in the data as used by Atkinson. Finally, 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 is clearly preferable. Such counts would have been available to Atkinson in the form of the original UPSID database [TODO: ref, link], from which the WALS database is a derivative. For our criticism here, we will use our own Phoible database (as introduced in Section 2).

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 consonant in WALS is minimally below 23 (Maddieson 2005a: 10) while the average number of vowels is just almost 6 (Maddieson 2005c: 14). Yet, in Atkinson's assessment of phoneme inventory size, the vowel inventory is given equal weight to consonant inventory, 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 by Atkinson. However, the number of tonal oppositions is almost always lower than the number of vowel oppositions (in Phoible there are three African languages listed that have more tones than vowels, viz. Metta, Bana and Babungo TODO: check!). In Phoible,

the average number of tonal oppositions is slightly lower than one per language, which means that Atkinson's assesments of phoneme inventory size are strongly biased toward tonal oppositions. Aggreviating this implicit weighting is that tonal oppositions show a strong geographical preference for Africa and South-East Asia, as can be immediately seen in the original WALS map (Maddieson 2005b).

By counting vowel inventory and tonal oppositions as independent characteristics Atkinson introduces yet another implicit weighting, because these two characteristics are actually positively correlated ($r = 0.33$, $p = 1.9\text{e-}14$ using WALS; $r = 0.21$, $p = 4.9\text{e-}15$ using Phoible). This somewhat surpising correlation is explicitly noted by Maddieson in WALS (Maddieson 2005b: 59). How exactly this correlation should be interpreted is still unclear, but it results in an even stronger emphasis of languages with tone in Atkinson's assessment of phoneme sizes.

The third problem with using the WALS data on vowels is that only vowel quality differences are counted. However, 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, nasalisation and diphtongization in WALS (Maddieson 2005c: 14). Further possibilities, though less frequently attested, are pharyngalization or glottalization. The Phoible database includes all such vowel oppositions as described for the world's languages. The vowel quality inventory as it is classified in WALS is still correlated with this more inclusive definition of phonemic vowel oppositions in Phoible, but the two are far from identical to each other ($r = 0.51$). The WALS counts of vowel qualities result in an average of 6 vowels per languages, while the average number of vowel oppositions in Phoible is 10 per language.

In summary, Atkinson's assessment of phoneme inventory size is only a rough approximation of the actual phoneme inventory sizes. His assessment is still correlated with the actual number of phonemic oppositions as included in the Phoible database ($r = 0.58$), but the two show clear differences. The differences are illustrated in Figure 1 and Figure 2, showing the distribution of phoneme inventory sizes over macroareas according to Atkinson and Phoible, respectively. The predominance of phoneme size in Africa is clearly an artifact of the specific way in which phoneme size is established by Atkinson.

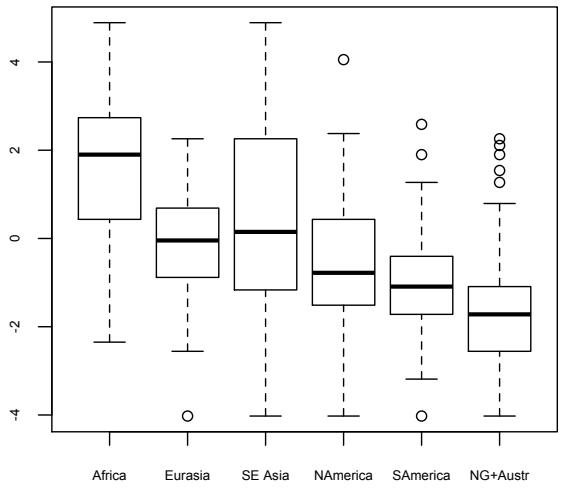

Figure 1. Atkinson's assessment of phoneme inventories classified according to areas. There appears to be a clear cline with Africa having most phonemes, followed by Eurasia and South-East Asia, with the Americas, New Guinea and Australia having less phonemes. The y-axis shows the sum of z-scores of the three WALS chapters used.

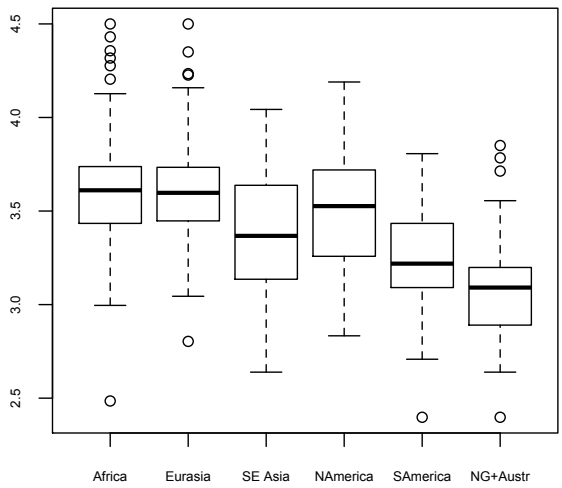

Figure 2. Phoible data on phoneme inventories classified according to the same areas. Africa and Eurasia still seem to be slightly ahead of the other areas, though North America is close, while South-East Asia is lowered relatively to these areas. The y-axis shows the logarithm of the actual number of phonemes as found in a language.

4. The impact of speaker community size

There is an observed positive correlation between the phoneme inventory size of a languages and the speaker community size (Trudgill 2004; Hay & Bauer 2007). This observation is reiterated by Atkinson, and we find it also using the Phoible database ($r = 0.30$, $p < 2.2\text{e-}16$). Note that for this correlation, we used the logarithm of population sizes and the logarithm of the phoneme inventory sizes. The analysis of the expected distribution of phoneme inventories is still not settled (Lehfeldt 1975; Justeson & Stephens 1984; Cysouw 2010), but using a logarithm seems to be better than taking the raw numbers. This correlation appears to be solid, though it is still far from clear how to explain it. However, for this paper we will simply accept the correlation as given, as 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 should be the population size that might have some kind of influence on language structure. It is 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 (where recent is defined as less than 10^4 years ago). Finally, the reason for a speaker population to grow large has various socio-political 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 close relative Y.

Given this perspective, the speaker population is a factor to account for in the measurement of inventory size. The more so as the geographical distribution of large speaker communities is not random at all. There is a strong bias of large speaker communities to occur in Africa and Eurasia. Figure 3 shows the geographical distribution of languages with more than 10^5 speakers in red, while all other languages in the Phoible database are shown in grey. The geographical bias is striking.

This geographical distribution of large speaker communities is influencing the geographical distribution of phoneme inventory sizes, favoring Africa as being a region with large phoneme systems. As the reason why Africa currently has many languages with large population sizes is surely not a factor related to the origin of human languages (i.e. it is *not* the case that there are many large languages in Africa because Africa is the point of origin of modern humans), this factor has to be

removed when the distribution of phoneme systems across the world's languages is investigated.

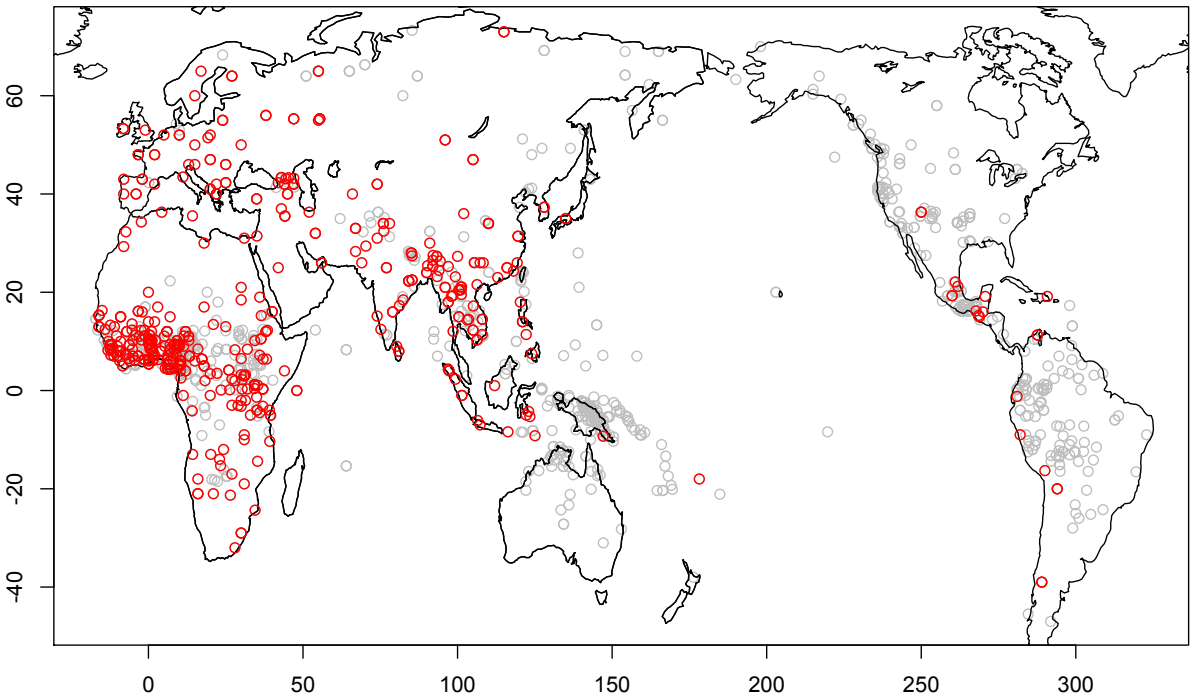

Figure 3. Geographic distribution of languages with more than 100.000 speakers, showing a strong preference for Africa, Europe, South Asia and South-East Asia.

Although it is not clear that there is a linear relationship between phoneme inventory size and speaker community size (contrary to this assumption there actually seems to be a major effect in the range of 10^4 to 10^5 speakers), we performed a simply linear regression of phoneme inventory size (in logarithms) to speaker community size (also in logarithm), and then investigated the residuals. The macroareal division of these residuals is shown in Figure 4. Comparing this distribution to the earlier distributions in Figure 1 and Figure 2, North America now is the area with the (relatively) largest phoneme inventories.

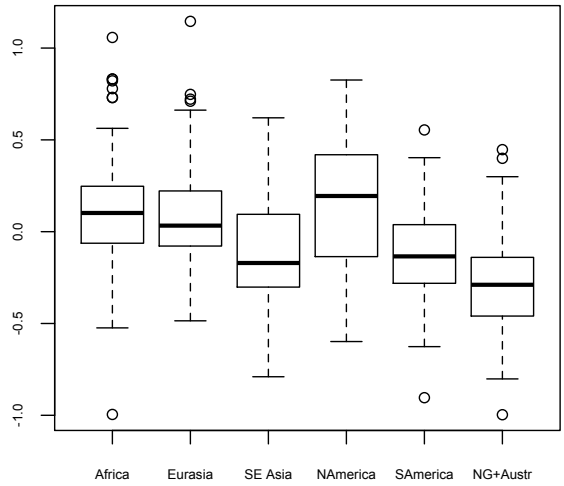

Figure 4. Residuals of phoneme inventory size after regression to population size. North America now is slightly ahead of Africa and Eurasia. The y-axis shows residuals after regression.

5. The geography of phoneme inventory size

Before trying to infer the optimal origin of current diversity in phoneme inventory size, we would just like to have a look at the approximate geographical distribution of phoneme inventories across the world. For the following maps, we use red for large phoneme inventories and blue for small phoneme inventories. Filled red dots signify the top 0.05 quantily and open red dots the top 0.25 quantily of sampled languages. Inversely, filled blue dots signify the bottom bottom 0.95 quantily and open red dots the bottom 0.75 quantile of the sampled languages.

First, Figure 5 shows the geographical distribution of the most extreme phoneme inventory sizes according to the assessment by Atkinson. There is a strong preference for large phoneme inventories to occur in Africa and South-East Asia, and this distribution is strongly reminiscent of the distribution of tone systems. That is no accident, as the phoneme inventory as determined by Atkinson is actually strongly influenced by having tone or not (see Section 3).

Second, Figure 6 shows the geographical distribution of the most extreme phoneme inventory sizes according to the Phoible database. Africa and South-East Asia still are ‘red’ areas, but Europe, the Caucasus and North America (areas long known for their often complex consonant systems) are also turning up. Finally, Figure 7 shows the residuals after regression to population size, which even more strengthens North America as being the centre of large phoneme inventories.

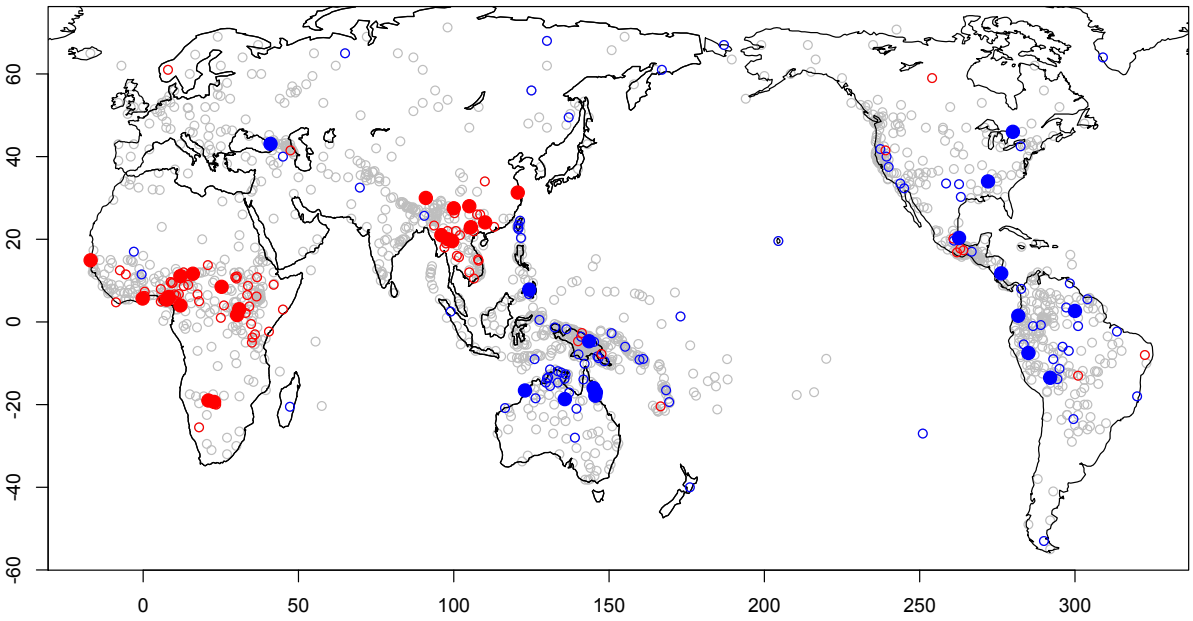

Figure 5. Inventory size according to Atkinson's assessment.

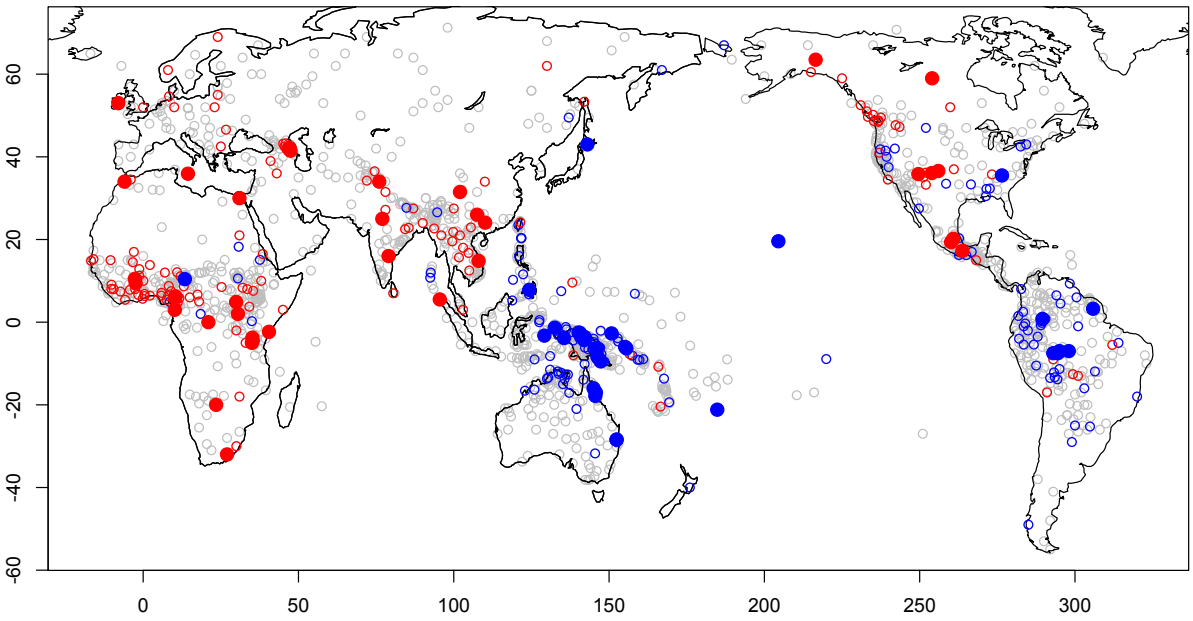

Figure 6. Inventory size according to the Phoible database.

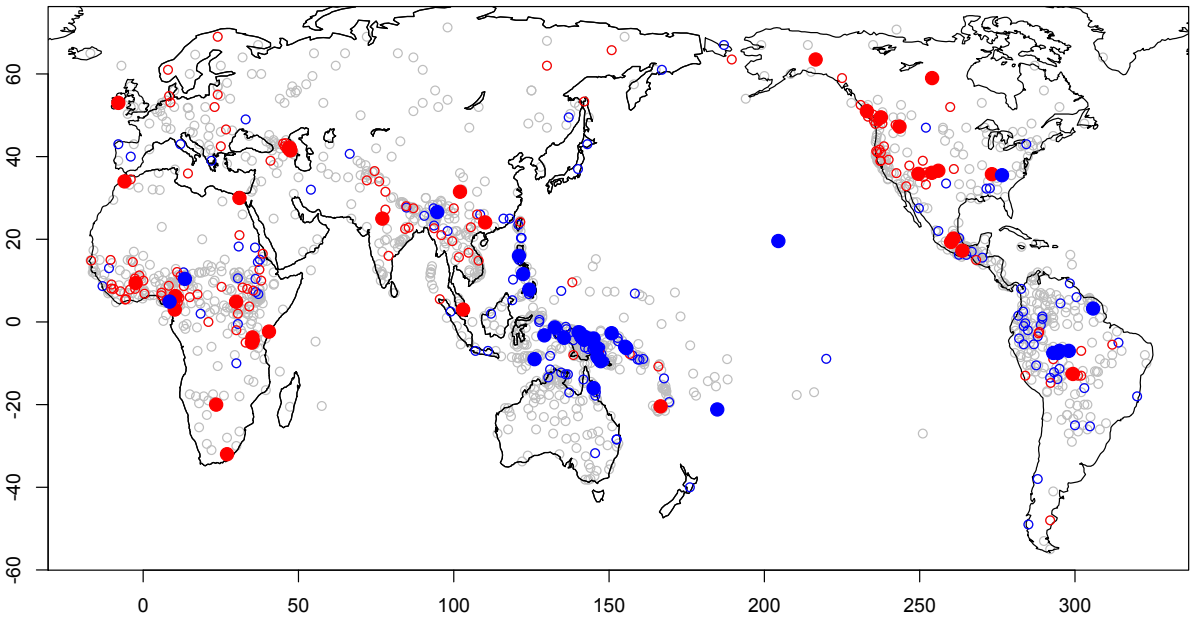

Figure 7. Phoible inventory size as residuals after regression to speaker community size.

6. Measuring geographical distance between languages

[TODO]

7. The internal structure of phoneme systems

[TODO]

8. Conclusion

References

{Bibliography}

Cysouw, Michael. 2010. On the probability distribution of typological frequencies, in: Christian Ebert, Gerhard Jäger & Jens Michaelis (eds.), *The Mathematics of Language*. 29-35. Springer: Berlin.

Hay, J & L Bauer. 2007. Phoneme inventory size and population size. *Language* 83(2). 388-400.

Justeson, John S & Laurence D Stephens. 1984. On the relationship between the numbers of vowels and consonants in phonological systems. *Linguistics* 22. 531-545.

Lehfeldt, Werner. 1975. Die Verteilung der Phonemanzahl in den natürlichen Sprachen. *Phonetica* 31. 274-287.

Maddieson, Ian. 2005a. Consonant inventories, in: Martin Haspelmath, Matthew S Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of Language Structures.* 10-13. Oxford University Press: Oxford.

Maddieson, Ian. 2005b. Tone, in: Martin Haspelmath, Matthew S Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of Language Structures.* 58-61. Oxford University Press: Oxford.

Maddieson, Ian. 2005c. Vowel quality inventory, in: Martin Haspelmath, Matthew S Dryer, David Gil & Bernard Comrie (eds.), *World Atlas of Language Structures.* 14-17. Oxford University Press: Oxford.

Martin Haspelmath, Matthew S Dryer, Bernard Comrie & David Gil (eds.).2005. *The World Atlas of Language Structures,* Oxford University Press: Oxford.

Trudgill, Peter. 2004. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305-320.