Appendix B: Creation of the Linguistic Heterogeneity Measure

Updated June 16, 2009

Our starting point in creating a measure of linguistic heterogeneity among a country's ancestor population is to determine the language spoken by people arriving from each potential source country. For each of the 165 countries in our matrix, we do our best to choose the dominant or most prevalent language of the country's population in the year 1500. Whenever possible, we use historical summaries to determine what was the largest ethnic group in the year 1500, and the language spoken by that group at that time. In some cases where historical information was not available, we use the current day (indigenous) language of the largest current indigenous group. Obviously our method is flawed in ignoring any heterogeneity of languages spoken within a source country. This problem is especially acute because our definition of "country" uses current borders, which are often unrelated to linguistic or cultural fault lines at the time that people emigrated. Thus, for example, immigrants from Sicily and Venice, who would not have been able to understand each other, are treated as having spoken the same language. However, as much of the heterogeneity that we measure relates to gross differences in language among the sources of a country's population (such as Amerindians vs. Europeans), our hope is this mis-measurement will not be too severe.

We then construct a matrix of linguistic distance among each pair of source country languages. The starting point for linguistic distance is a tree showing the relations among all current and known past languages (Gordon, 2005). Every language can be characterized by its family (such as Indo-European or Uralic), and then a series of "nodes," representing the branching points of the language tree, ending in the language itself. For example, the full tree of Spanish is Indo-European, Italic, Romance, Italo-Western, Western, Gallo-Iberian, Ibero-Romance, West Iberian, Castilian, Spanish. Any two languages in the same family can be connected by going up and then down a certain number of nodes. For example, the tree for Italian is common with Spanish through Italo-Western, and is then followed by Italo-Dalmatian, Italian. Italian and Spanish thus have four nodes in common. We measure the distance between any pair of languages as[1]

$$d_{i,j} = 1 - \left(\frac{\text{\# of common nodes between } i \text{ and } j}{\frac{1}{2} \times \left(\text{\# of nodes for langauge } i + \text{\# of nodes for language } j \right)} \right)^{\lambda} .$$

Languages from different families have no nodes in common, and so the distance between them is one. The parameter λ is assumed to be between zero and one, implying

[1] The only difference between our method and Fearon (2003) is that in the denominator he uses 15, which is the maximum number of nodes for any language.

that earlier common nodes have a larger weight in the distance function than later ones. In practice, we follow Fearon (2003) in assuming $\lambda = 0.5$.[2]

Finally, we combine our linguistic distance measure with the information on source countries in the matrix. Let ***L*** be the matrix of linguistic distances and ***A*** be the matrix with current countries as rows and source countries as columns. Our new measure, which we call "historical linguistic heterogeneity," is the diagonal of ***ALA'***.

Information on which indigenous language we chose for each country in 1500, and why we chose them, can be found in three Excel spreadsheets that make up the rest of this appendix (These spreadsheets were prepared by Joshua Wilde). The first spreadsheet, Language Spreadsheet.xls, contains each country and selected language, along with the 15 branch classification used to calculate linguistic distance. It also contains a source column which indicates whether this language was chosen based on data from Fearon (2003) or Ethnologue (Gordon, 2005), or both. Two more excel spreadsheets, Fearon Notes.xls and Ethnologue Notes.xls contain detailed explanations as to how we interpreted the data and made our final decisions on which language would be assigned to each country. The source information tab in Language Spreadsheet.xls gives further instructions on how to use these two spreadsheets, which countries are contained in each, and how to read them.

References

Fearon, James, D., "Ethnic Structure and Cultural Diversity by Country," *Journal of Economic Growth*, 8:2, June 2003, 195-222.

Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/

[2] Experimenting with value of lambda in the range 0.25-0.75 had very little effect on the results shown below.