Empirical Seminar: Draft Paper

Global Inequality, Linguistic Distance, and Diffusion

Kyle Chan

May 25, 2016

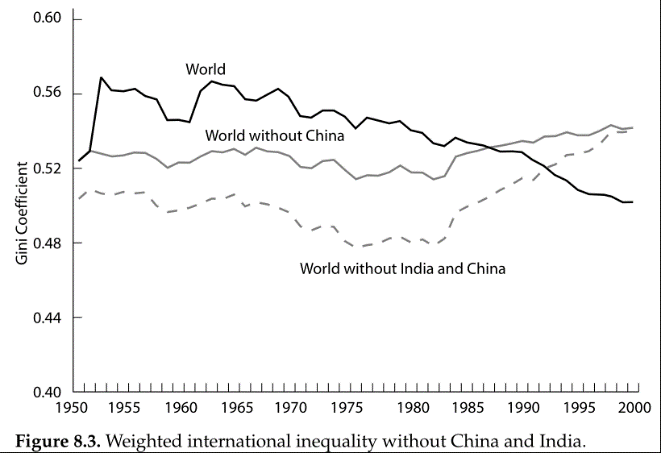
**Abstract**

Why does global inequality exist, and what explains patterns of economic development around the world today? Prior research has shown strong relationships between supranational factors such as colonial legacy and geography on the one hand and current levels of development on the other. This study examines the relationship between one such supranational factor—linguistic distance, as a proxy for cultural distance more generally—and patterns of economic growth over time in the latter half of the twentieth century. After controlling for geography, income levels, and trade, linguistic distances between countries is found to be significantly associated with similarities in patterns of economic growth. The transnational diffusion of ideas and practices—i.e., technology—is posited as a key mediating process: technology is more likely to spread between countries that are more linguistically—and culturally—proximate, leading to differing economic growth trajectories.

Global income inequality is massive and persistent. While global income inequality appears to have declined in recent decades, this is primarily due to the economic rise of China (Firebaugh and Goesling 2004; Milanovic 2005). Excluding China from the analysis, the global population-weighted Gini coefficient remains above 0.52 and has in fact grown in recent decades (Milanovic 2005). Within-country inequality has recently become more salient in national politics and in social science research, driven in no small part by the work of Thomas Piketty (2013). However, between-country inequality remains the predominant driver behind global inequality today (Bourguignon and Morrisson 2002; Ravallion 2014; World Bank 2005) and deserves further investigation alongside within-country inequality.

What explains global income inequality across countries? Many conventional approaches to the study of development emphasize the characteristics of individual countries. Most prominently, the Solow growth model, named after its inventor Robert Solow (1956) and known within economics as the neoclassical growth model, has long dominated work on economic development among researchers and policymakers. While the Solow growth model makes no explicit assumptions about the nature of inter-country relationships, its design tends to emphasize individual country characteristics such as national labor supply, capital stock, and level of technological progress.

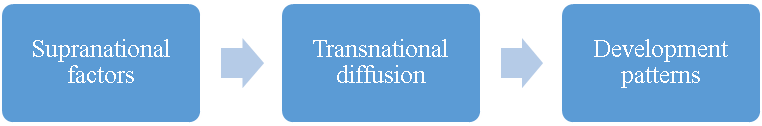
In contrast to this approach, this paper joins a broader shift in research focus within sociology and the social sciences away from the characteristics of individual countries and toward an examination of supranational factors and their effects on relationships among countries. Dependency theory and world-systems theory, as pioneered by Wallerstein (1974) and others, utilizes a Marxist framework to understand a persistent power relationship between so-called “core” and “periphery” countries. World-polity theory, as developed by John Meyer and others (Meyer, Boli-Bennett, and Chase-Dunn 1975; Meyer 1980; Meyer et al. 1997), builds on a Weberian and new institutionalist emphasis on culture to explain the spread of certain institutional forms around the world. The enduring effects of colonial legacy on contemporary patterns of development has become the focus of a rapidly growing body of quantitative research that has identified important differences in colonizer identity, colonial regime structures, and settlement patterns (Acemoglu, Johnson, and Robinson 2001, 2002; M. K. Lange 2004; M. Lange, Mahoney, and Hau 2006). The effect of war and inter-state competition on the formation of nation-states and the motivations of political leaders to promote economic growth have been explored in detail by Centeno (2002), Tilly (1990), and many other scholars. Other supranational patterns have also been explored, driven by factors such as geographical proximity, religion and ideology (including communism), and so on.



**Figure 1**. Population-weighted Gini coefficients over time with and without China and India. Directly from Milanovic (2005).

**Transnational Diffusion**

One hypothesized mechanism by which supranational factors affect patterns of development across groups of countries is that of transnational diffusion. In the context of economic development, the diffusion of technologies—from physical technologies such as computers and blast furnaces to organizational technologies such as double-entry bookkeeping and professional curricula vitae—across countries has been uneven in terms of historical timing and levels of adoption. The levels of adoption and types of utilization of certain technologies may in turn affect a country’s prospects for economic development.



Reduced form

**Figure 2**. Supranational factors may influence patterns of transnational diffusion, which in turn may affect development patterns across groups of countries. This paper examines the “reduced form” relationship between one supranational factor—linguistic distance as a proxy for cultural distance—and patterns of economic growth.

Historically, certain networks or channels of technological diffusion have been particularly well explored. The so-called “Republic of Letters” across Western Europe from the early modern period through the Scientific Revolution and the Enlightenment was comprised of formal and informal exchanges of ideas, often by handwritten letter, among individual thinkers and later scientific societies. Leibniz alone had a network of over 400 correspondents across France, Italy, England and even China (via the Jesuits) through which he shared and discussed new developments in mathematics and natural philosophy (Harris 2006). During the late Industrial Revolution, shortly after the Gramme dynamo, the first commercially viable electric power generator, was invented in Belgium around 1871, an American physics professor at Cornell built a working replica using accounts published in the British journal *Engineering* (Brittain 1974). In the 1880s as Meiji Japan strove to rapidly modernize, Japanese cotton-spinning firms imported British mill designs, engineers, and building materials to Japan wholesale, sometimes down to the very bricks from Lancashire (Saxonhouse 1974). A similar story occurred later in the 1970s when the South Korean conglomerate Daewoo invited entrepreneurs from Bangladesh to visit South Korean clothing factories and learn the textile trade. Bangladesh’s textile exports are now the second largest in the world after China’s. (South Korea, in turn, received industrial technology and know-how from the Japanese during a half-century period of colonization.)

While this paper does not directly test for the effects of transnational diffusion as a mechanism *per se* (which Wejnert (2005) does for diffusion and democracy), this paper takes a step toward exploring the relationship between broader supranational factors that may influence diffusion networks and patterns of development over time. In particular, this paper investigates the relationship between linguistic distance—how similar are the dominant languages of any two countries?—and patterns of development. While sharing similar languages may simplify the task of translation to some extent, this paper does not consider the direct effect of language on technological diffusion *per se* but rather uses linguistic distance as a proxy for broader forms of cultural distance that may in turn influence the formation of diffusion networks.

A broad body of theoretical and empirical literature has emerged across sociology and the social sciences that investigates the relationship between culture and transnational diffusion. One set of literature within sociology has focused on the institutional sources of various patterns of transnational diffusion. In their theoretical discussion, Strang and Meyer (1993) stress the potential role of cultural linkages and processes of theorization by which certain causal chains are abstracted from their original context and thus become more portable across cultural contexts. Other work by Kaufman and Patterson (2005) has focused on tracing the particular channels and networks of diffusion for cultural practices such as the game of cricket. A growing set of empirical studies of policy diffusion have identified key mechanisms such as shared “cognitive-psychological frameworks” that allow policies such as pension privatization to spread more quickly across certain sets of countries (Weyland 2005). The diffusion of democracy in particular has been the main focus of a number of studies that use large country-year datasets to identify specific causal factors that influence the likelihood and timing of democratization around the world (Wejnert 2005; Weyland 2012). Within economics, a body of literature has emerged focusing on the specific technologies that are diffused and their effects on economic development (Comin, Hobijn, and Rovito 2008; Comin and Hobijn 2010; Keller 2004; Spolaore and Wacziarg 2009)

This paper contributes to the literature on culture and transnational diffusion by identifying a novel relationship between cultural similarity (as proxied by linguistic distance) and patterns of economic growth. The use of linguistic distance as a proxy for cultural distance may offer a new source for uncovering the roots of certain transnational diffusion networks.

In addition, this paper addresses a number of theoretical and methodological issues that have arisen in the study of supranational patterns of economic development that will be discussed in the following section.

**Strict Causality vs. Patterns of Association in Cross-Country Studies**

This paper advocates a shift away from an exclusive focus on strict causal models that attempt to isolate and identify the effects of individual factors on patterns of economic development. There are two strongly interrelated reasons for this. First, given the limited sample size of any country-level analysis (there are currently 193 United Nations member states) relative to the myriad potential explanatory factors, the system is statistically overdetermined, even with country-year data.

Second, there exists too much endogeneity among potentially important explanatory factors to be able to sufficiently isolate a single factor and identify its causal effects. For example, in the debate over whether institutions or capital accumulation (including human capital) better explain variation in development outcomes, institutions and (human) capital are too intimately interrelated to make such a distinction useful. Greater levels of education likely strengthen institutions, and greater institutional capacity likely increases levels of education. Rather than attempt to cut the Gordian knot of causal identification in country-level patterns of development, I argue that researchers should instead devote more attention to mapping out cross-country patterns and to examining the specific channels or mechanisms by which these patterns may arise.

**Levels of Income versus Patterns of Growth**

Previous work on economic development has focused on income levels, which suffer from high temporal autocorrelation. A number of studies have shown the long-term persistence of relative income levels over time (Chanda, Justin Cook, and Putterman 2014; Comin, Easterly, and Gong 2010; Hibbs and Olsson 2004; Putterman and Weil 2010). Moreover, the historical mechanisms of accumulating returns and path dependence may partly explain why there exist such high degrees of temporal autocorrelation over the longue durée. Thus, recent studies attempting to explain contemporary variation in levels of development may simply be detecting earlier historical income levels rather than the effects of more contemporary diffusion networks or other supranational development factors.

To address this problem, this paper examines patterns of economic growth rather than levels of income. While the relative rankings of countries by income may be fairly stable over time, it is less clear that the specific temporal shapes of their economic growth rates over time would be contingent on their prior levels of income. Moreover, I have taken the additional step of standardizing growth rates: for any given country, its annual growth rates have been de-meaned and re-scaled to unit standard deviations. This further removes the effects of prior levels of income on patterns of economic growth. Without this step, one could expect a meaningful relationship between higher levels of income and lower average growth rates (following the beta-convergence of the neoclassical growth model) and with lower variance due to lower volatility in growth rates (Barro and Sala-i-Martin 1992; Mankiw, Romer, and Weil 1992). As a final step, I have included current income levels as a control variable in my analysis.

An examination of patterns of economic growth is not only important from a methodological standpoint; understanding variation in patterns of development is an important sociological phenomenon to investigate in itself. Static income levels capture only one dimension of development, ignoring the important ways in which the experience of development as an evolving process over time differs across countries.

A number of scholars of development have examined the ways in which groups of countries have differed in their historical paths. Peter Evans (1995), Atul Kohli (2004), and Arthur Sandbrook et al. (2007) have charted various pathways of development experience outside the West from the state-led, export-oriented experiences of a number of East Asian countries to the social democratic political economies of Kerala and Mauritius. Any student of communism, particularly of the Soviet Union and Communist China, is familiar with the peculiar focus on heavy industry and economic autarky of these countries’ development experiences. Within Latin America, the United Nations’ Economic Commission for Latin America (ECLA) strongly influenced the import-substitution industrial policies of many countries in the region during the twentieth century (R. R. Kaufman 1990). Finally, as scholars such as Frank Dobbin (1994) have shown, even among Western capitalist there may be drastically different historical paths of development from the United States’ decentralized, local-level approach to France’s centralized administrative model.

**Rethinking Geography and Regional Categories**

Many scholars, policymakers, and private sector actors take certain regional groupings of countries for granted. For example, multinational corporations often break down global markets into regional categories such as North America, Latin America, Europe, Middle East, Africa, and Asia Pacific. While this may be useful heuristically and may be sufficiently valid for certain purposes, not enough work has been done to question the validity of this geographical assumption. Certain examples of exceptions to geographical factors are especially prominent.

The Anglophone countries are perhaps the best example where the commonalities across Britain, the United States, Canada, Australia, and New Zealand stand in direct contradiction to their vast geographic dispersion. In terms of levels of income, these countries are much more closely related than they are to their geographical neighbors. Moreover, on a large number of economic, political, social, and cultural dimensions, these countries constitute a fairly coherent grouping that defies their relative geographical positioning. In terms of models of social service provision, for example, these countries tend to stand apart from other industrialized countries for their relatively “commoditized” welfare systems (Esping-Andersen 1990).

Geography also breaks down as a useful predictor of sociopolitical characteristics when there are sharp breaks in this characteristics despite relatively continuous geographic positioning. While countries across Western and Eastern Europe are distributed on a fairly continuous basis geographically, there exists a sharp break across the former Iron Curtain that appears in contemporary economic, political, and social measures. Such invisible boundaries also may exist between Northeast Asia and Southeast Asia as well as between Northern Africa and Sub-Saharan Africa.

To address this problem, I examine linguistic distance while controlling for geographic distance between countries. As I will later demonstrate, there are interesting relationships in patterns of economic growth that defy geographic proximity as a dominant factor.

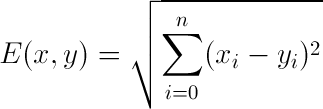
In addition to these contributions to theory, this paper also makes data and methodological contributions. In terms of data, this paper employs a high-quality state-of-the-art linguistic dataset based on the presence or absence of specific phonemes that allows for more continuous measures of linguistic distance, particularly over the standard Ethnologue dataset. In terms of methodology, this paper introduces a non-parametric statistical procedure commonly used in population genetics research that would be especially useful to sociologists given its focus on relationships between data points rather than the characteristics of individual observations. These two additional contributions will be discussed in greater detail in the next section.

**Data and Methods**

Data for this study come from publicly available cross-country datasets. For each variable, distance matrices are generated which contain distance measures for every pair of countries, yielding a symmetric *n* x *n* matrix containing distinct distance values where *n* is the number of countries in the sample. Simple and partial Mantel tests, commonly used in population genetics, are used to calculate correlation coefficients between these distance matrices and corresponding significance levels. Conventional statistical techniques such as OLS regression cannot be used due to complete autocorrelation within each distance matrix. The primary dependent variable is differences in patterns of economic growth and the primary independent variable is linguistic distance. Control variables include geographic distance, differences in income levels, and bilateral trade.

*Patterns of Economic Growth*

Annual per capita GDP data by country from 1950 to 2008 come from the Maddison Project (Bolt and van Zanden 2014). This timeframe was selected in order to maximize data availability and to use conventional start (1950) and end (latest available year: 2008) years that are likely relatively orthogonal to the variables of interest. National per capita GDP values are presented in the original dataset in constant 1990 international Geary-Khamis dollars, which are adjusted for purchasing power (PPP). From this, I have calculated rolling annualized five-year averages, yielding one economic growth rate value for each country-year (trailing five years) from 1955 to 2008. The use of five-year average annualized growth rates reduces the effect of measurement error from any single year. These annualized five-year average economic growth rates are then standardized within each country across all years such that each country’s mean annualized growth rate is zero with unit standard deviation. Standardizing the economic growth values within countries in this way reduces the effect of historical income levels, which are known to be associated with average growth rates and growth rate volatility as discussed earlier.



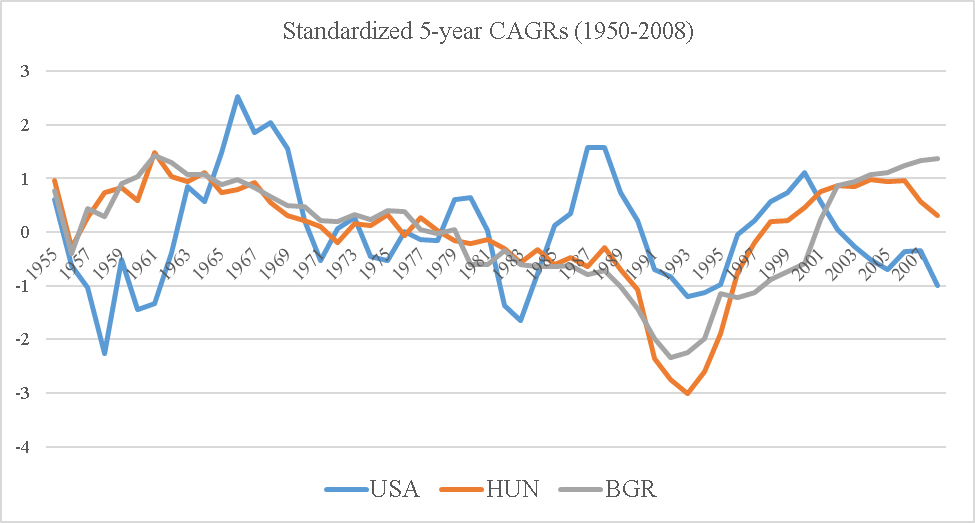
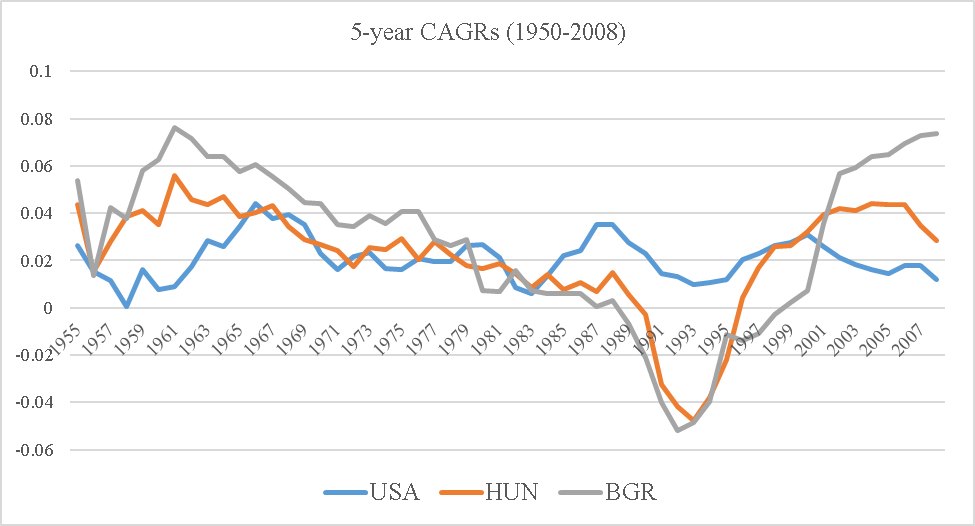
**Equation 1**. Euclidean distance formula between any two points *x* and *y* for *n* dimension

These standardized values for each country-year are then used to generate a distance matrix of growth pattern distance values for every country-pair based on the Euclidean distance across all years for any two countries. In the Equation 1, *x* and *y* are two different countries and *i* is each year’s five-year annualized growth rate for *n* years 1950 to 2008. Figure 3 shows growth rates and standardized growth rates for the United States, Hungary, and Bulgaria. In both charts, the shapes of Hungary and Bulgaria’s growth rates are much more similar to each other than to the shape of U.S. growth over the same period.

*Linguistic Distance*

While previous work has examined the relationship between linguistic distance and economic development, much of this work has relied on overly coarse linguistic data. The Ethnologue dataset of world languages is widely used in the social sciences, particularly within economics and political science (Fearon 2003; Simons and Gordon  R.G. 2006). It was created by linguistic experts who manually coded and produced a linguistic tree for all known languages. However, the only way to derive the linguistic distance between two languages consists of counting the number of common nodes they share. This approach yields linguistic distance measures that may be coarser than other distance measures such as genetic distance. Thus, when Spolaore and Wacziarg (2009) perform a multivariate regression with both linguistic distance using the Ethnologue data and genetic distance as right-hand-size variables, they find genetic distance to be a better predictor of economic development. This is more likely due to the fact that the linguistic distance measure was coarser than the genetic distance measure, an issue they note.

To address this problem, I use a state-of-the-art linguistic dataset of 2,082 languages with binary presence/absence data for 728 phonemes. This publically available dataset is taken directly from a version of the Ruhlen database that was modified and used by Creanza et al. in their 2015 paper in the



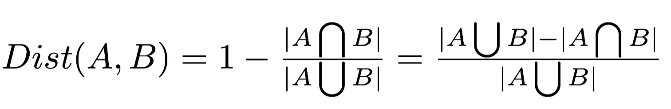
**Figure 3**. Non-Standardized and Standardized Annualized Growth Rates for Three Countries

The three countries shown are the United States, Hungary, and Bulgaria. The y-axes of both graphs are in percent annualized growth rates. Years begin in 1955 due to the use of five-year trailing averages.

*Proceedings of the National Academy of Sciences*. The primary advantage of this dataset is that one can use it to derive more continuous linguistic distance measures than the Ethnologue data.

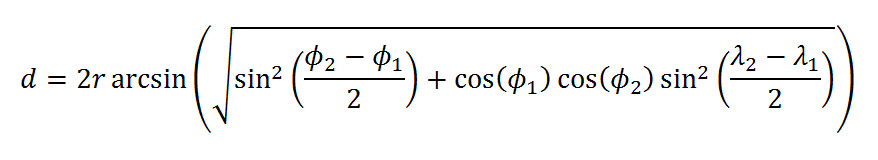
From the full set of 2,082 languages, each country is assigned a single “dominant” language corresponding to the language within that country with the greatest number of speakers. The assignment of a single language to each country is a useful simplification but raises a number of issues worth noting. For some countries, the most commonly spoken language may in fact constitute a relatively small fraction of the total population due to the existence of other widely spoken languages. The assignment of a single “dominant” language is a simplification that necessarily excludes some linguistic information. Ultimately, this simplification should make this paper’s findings more conservative given that the use of a single “dominant” language is a less precise measure of cultural similarity.

**Equation 2**. Jaccard distance formula between any two points *A* and *B* based on phonemic presence/absence data



In addition to this methodological issue, at a more data-specific level, there is an issue with 15 (of 115) countries where the “dominant” language as determined by plurality of speakers in the Ruhlen dataset does not correspond to the true plurality-spoken language according to external sources. As a result, a number of manual substitutions were made based on the CIA World Factbook data on most widely-spoken language by country. Most of these edits occurred in English- and Spanish-speaking countries, and the closest language approximation was used. Ultimately, the use of non-edited versus edited linguistic datasets did not change the findings (See Appendix A for a list of manual edits to the linguistic data.)

**Equation 3**. Haversine formula for calculating the great-circle distance between any two latitude-longitude pairs.



Using this new dataset where each country is assigned a single “dominant” language, I created a linguistic distance matrix where the linguistic distance between any two countries is calculated as the Jaccard distance on the presence/absence data for each phoneme. The Jaccard distance measure is commonly used within linguistics (Jaccard 1912). The difference between the number of phonemes present in either language and the number of phonemes present in only one language is calculated to yield the number of phonemes that are present in both languages. Then this is divided by the number of phonemes present in either language in order to rescale for absolute number of phonemes. This last step is necessary to adjust for languages that have a large number of phonemes.

*Geographic Distance*

Geographic data for each country comes from Google’s online set of canonical public-use databases for software developers (https://developers.google.com/public-data/docs/canonical/countries\_csv). For each country, geographical data consist of latitude and longitude values for the country’s geographic center. The geographic distance between any two countries is calculated as the great-circle distance between these two sets of latitude-longitude coordinates using the haversine formula. This provides the shortest distance between two points on the surface of a sphere. All country-pair geographic distances are then logged to adjust for scale.

*Income Levels*

2008 per capita GDP income levels come from the Maddison Project. As with the economic growth data, these per capita GDP values are in constant 1990 international Geary-Khamis dollars, which are adjusted for purchasing power (PPP). A distance matrix for income levels is created using the logged absolute difference between the 2008 per capita GDP income levels for each country-pair.

*Bilateral Trade*

Bilateral trade data comes from UN Comtrade, which is the United Nations’ publically-available online database of official global trade statistics (http://comtrade.un.org/). Derived from national import records, the UN Comtrade dataset provides a monetary value of one-way trade flows in nominal U.S. dollars for each importing country by exporter source. For example, in 2000 the U.S. imported $233 million in goods and services from exporting country Canada.

Rates of missing data are much higher for these bilateral trade data than for other distance measures. In order to increase the number of available observations and to reduce the particular effect of any single year, for each import-export country-pair the annual trade values over the decade 1991-2000 were averaged wherever data was available. The final number of distance values and proportion of data that was missing will be discussed later in the paper.

Both directions of trade flows for every country-pair were summed to produce a single total trade value for each country-pair. For example, U.S. exports to Canada and Canadian exports to the U.S. are summed to produce a single bilateral trade value for the U.S.-Canada country-pair.

**Simple and Partial Mantel Tests**

Simple and partial Mantel tests are commonly used methodological tools in population genetics to evaluate relationships between distance matrices (P. Legendre 1993; Mantel 1967; Sokal 1979). Conventional regression methods cannot be used effectively on distance matrices due to the complete autocorrelation inherent within any distance matrix. Within the standard regression framework, autocorrelation within sub-sample groupings is often addressed through the use of clustered standard errors. However, for distance matrices, autocorrelation does not exist at any sub-sample level but rather permeates the entire sample due to the triangle inequality property: for any three points A, B, and C in the underlying raw data, their three corresponding distance values A-B, B-C, and A-C in the distance matrix will be mutually correlated by transitivity. Thus, conventional regression methods such as OLS will overestimate the significance of any relationship by failing to adequately compensate for this inherent correlation across distance values.

The Mantel test offers one approach to addressing this autocorrelation problem. By randomly permutating rows (or columns) of one of the distance matrices DX and then calculating correlation coefficients between the randomized distance matrix DX and the comparison distance matrix DY, one can computationally generate random null hypothesis correlation coefficients.[[1]](#footnote-1) Given a number of random permutations, a *p*-value is derived by checking how many of these randomized correlation coefficients exceed the value of the original correlation coefficient. For example, if 99,999 random permutations were performed and 9 of them produced higher correlation coefficients than the correlation coefficient being tested, the P-value would therefore be 9 plus 1 (for the original combination) out of 99,999 plus 1, in other words P = 10/100,000 or 1e104.

While the simple Mantel test as described above tests for the relationship between two distance matrices, the partial Mantel test allows for the assessment of such a relationship while controlling for one or more additional distance matrices. To evaluate the relationship between two distance matrices DX and DY while controlling for a third distance matrix DZ, the residuals from the partial correlations of DX on DZ and DY on DZ are calculated, yielding two new residual matrices DXZ and DYZ, respectively. Then the simple Mantel test with randomized permutations is performed on these residual matrices to generate a correlation coefficient and *p*-value that excludes the effect of the joint relationship of DZ on DX and DY.

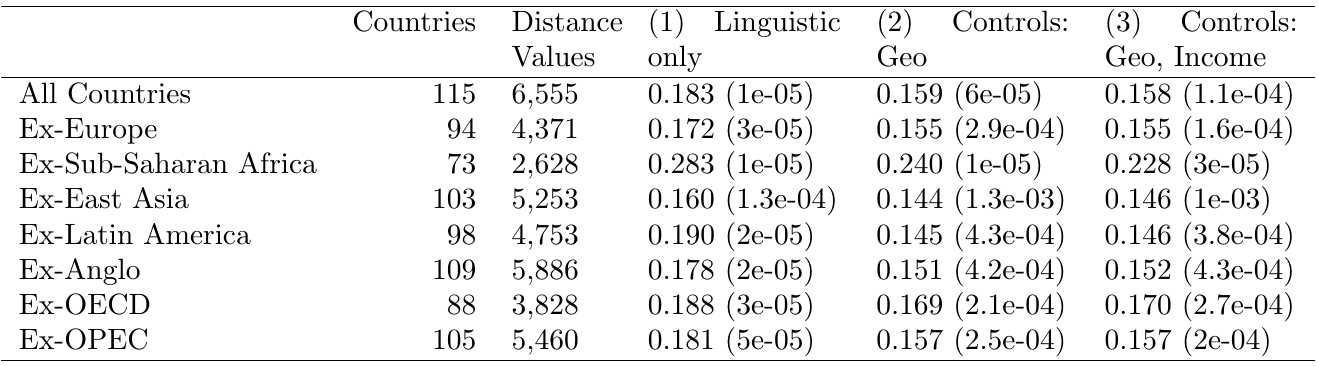
In the application of the Mantel test to the distance matrices in this analysis, the number of randomized permutations used to generate *p*-values is 99,999. The actual correlation coefficient being evaluated is the standard Pearson product-moment correlation coefficient, also known as Pearson’s *r*.

**Results**

There is a significant positive correlation between linguistic distance and growth pattern distance, even after controlling for geographic distance, income levels, and bilateral trade. This relationship is robust to the exclusion of certain subsets of countries that would likely significantly contribute to the relationship. Table 1 shows a summary of the correlation coefficients and their associated *p*-values across each set of controls and sample subsets (Appendix B includes a list of member countries for each subset). An analysis of trade controls is shown in a later section due to the higher rates of missing data.

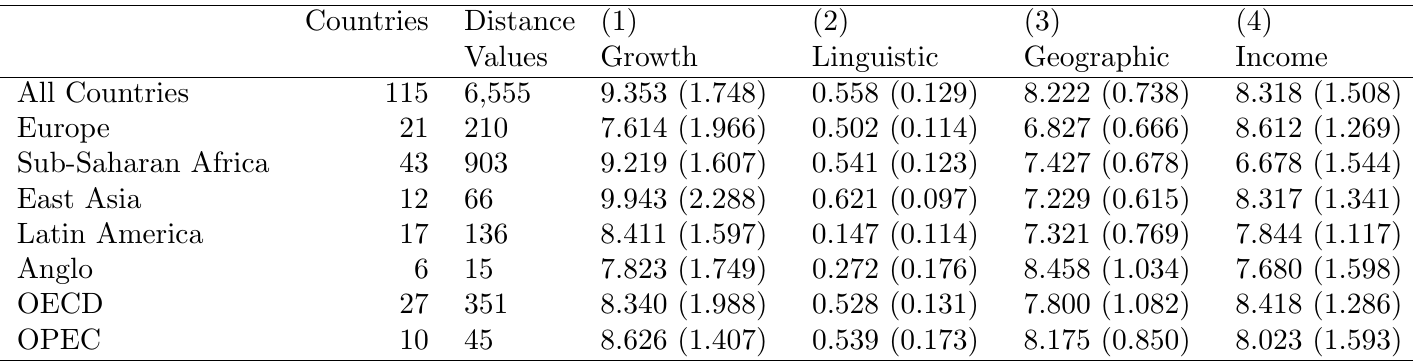
**Table 1**. Correlations between Linguistic and Growth Pattern Distances by Country Subset

Columns (1)-(3) contain Mantel *r* correlation coefficients with two-sided *p*-values in parentheses. Column (1) is a simple Mantel test between linguistic distance and growth pattern distance without controls. Column (2) and (3) add controls for log-geographic distance and both log-geographic distance and log-2008 income distance, respectively. All correlation coefficients are significant at the *p* < .05 level.



Columns (1) and (2) are distance values with no interpretable units. Column (3) contains log-geographic distance in log-miles Column (4) contains log-2008 income level distance values in log-2008 U.S. dollars. Standard deviations are in parentheses.

**Table 2**. Mean Distance Values by Distance Measure and by Country Subset



By comparing the growth-language correlation coefficients in column (3), which control for geography and income, we can see that while no country subset dominates the overall growth-language relationship, the exclusion of East Asian countries and the exclusion of Latin American countries reduces the overall correlation by the greatest amount. This implies that of the country subsets tested for in this analysis, the most significant drivers of the growth-language relationship are East Asian countries and Latin American countries.

Table 2 shows the average distance values with standard deviations for each country subset. From this, we can see that European countries and Anglophone countries have the closest growth patterns while Sub-Saharan African countries and East Asian countries exhibit the highest divergence in growth patterns. As expected, linguistic distances are smallest within Anglophone countries and within Latin American countries. These results shed additional light on which country subsets may be most significantly influencing the global growth-language relationship. Latin America’s close linguistic grouping along with its closer than average growth pattern clustering may explain how these countries as a group relative to the rest of the world may be driving the growth-language relationship. In contrast, East Asian countries, which constitute a particularly small country subset, may be contributing to the growth-language relationship due to wide within-subset variation.

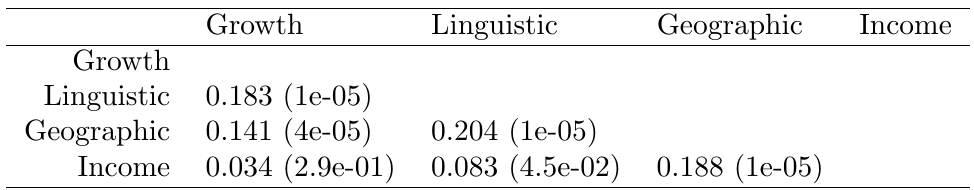
Within country subsets, however, there does not exist a significant correlation between growth patterns and linguistic distance. This implies that most of the overall relationship is driven by variations across country subsets rather than within country subsets.

*Correlation Table across Distance Variables*

Table 3 shows the correlation coefficients (and corresponding *p*-values via Mantel test) between each set of distance measures. These results offer a test of validity for the data and methods as well as several insights into the relationships across these factors.

First, the relationship between distances for growth patterns and income levels is close to zero and not significant at the 95% level. This validates the approach taken to insulate growth rates from income levels via within-country growth rate standardization.

Includes all countries (number of countries = 113; number of distance values = 6,555). *P*-values are in parentheses.



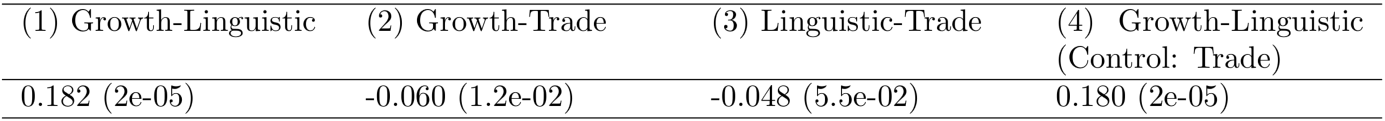
**Table 3**. Pairwise Correlation Coefficients between Distance Measures

Second, the relationship between geographic distance and linguistic distance as well as the relationship between geographic distance and income level distance is relatively strong and statistically significant. This reconfirms prior findings of the underlying effect of geographic distance on linguistic distance and on income levels.

Third, linguistic distance is more strongly related to growth patterns than is geographic distance. This implies that language may be a more meaningful way of understanding cross-country clustering patterns in development than geography or simple regional groupings.

**Table 4**. Correlation Coefficients with Bilateral Trade

Includes all countries (number of countries = 108; number of distance values = 2,373). *P*-values are in parentheses.



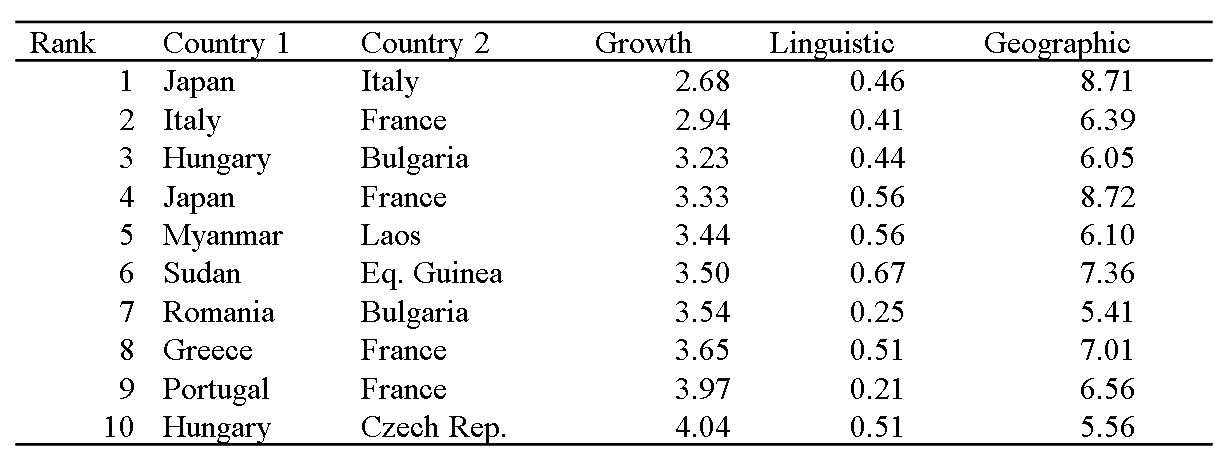
*Robustness to Trade*

The significant positive relationship between growth pattern distance and linguistic distance continues to hold, even after controlling for bilateral trade value. Table 4 shows the correlation coefficients and *p*-values for these simple and partial Mantel tests. Due to the relatively high proportion of missing data for bilateral trade value, trade cannot be controlled for alongside other control variables such as geographic distance. A separate algorithm is required to selectively omit all distance value observations that are not present in every distance matrix in the analysis. The total number of countries that can be included in the trade analysis is consequently restricted to 108 and the total number of distance values that can be used is 2,373, which is 47% out of 108 (108-1) / 2 = 5,778 total possible distance values.

In Table 4, column (4) shows the correlation coefficient between growth patterns and linguistic distance to be significant and positive, even after controlling for trade. Column (1) shows that the original simple correlation between growth patterns and linguistic distance holds, even when restricted to the country-pairs that are available in the limited trade data. As expected, columns (2) and (3) show that trade is significantly correlated with growth patterns and with linguistic distance. The negative correlation coefficient is due to the fact that the trade “distances” are in fact proximity rather than distance measures: larger bilateral trade values between any two countries indicates greater trade proximity.

Geographic distance is in logged miles.

**Table 4**. Top Ten Most Similar Country-Pairs by Growth Pattern



*Patterns of Growth Similarity*

Finally, a brief investigation of certain country-pair relationships yields several interesting findings. Table 4 shows the top ten most similar pairs of countries as measured by their growth patterns. The post-WWII economic boom countries in Western Europe and Japan dominate the list. Alongside this, several pairs of Eastern Bloc countries also exhibit high relative similarity in growth patterns. While proximity in language and geography tend to characterize the growth pattern clustering of these country-pairs, it is interesting to note broader historical factors that cut across these dimensions, as in the case of Japan and many Western European countries.

**Discussion**

The robust relationship between linguistic distance and patterns of economic growth lends empirical support to the broader notion that patterns of economic growth across countries may be strongly related to underlying supranational factors. While the relationship between factors such as geography, religion and ideology, communism, and colonialism on the one hand and levels of national income has been well documented, a focus on studying economic growth patterns in particular and their relationship with supranational factors such as language may yield further insights into the clustering of development experiences across countries.

As these clustering patterns become more clearly identified, an important next step would be to analyze the diffusion channels and networks that mediate these relationships. This would likely require the collaborative use of a wide range of methods from detailed comparative historical work to quantitative network analysis.

Taken together, these sets of research projects may offer a better understanding of how current development patterns came about and why global inequality remains so high despite the supposed availability of technology in the information age.

1. Note: There has been some controversy over the extent to which these randomized permutations produce an appropriate set of null hypotheses that sufficiently controls for spatial autocorrelation (Guillot and Rousset 2013; Pierre Legendre, Fortin, and Borcard 2015). [↑](#footnote-ref-1)