Population Relatedness and Cross-Country Idea Flows: Evidence from Book Translations

Andrew Dickens[†]

Final Version

This paper establishes a robust relationship between idea flows across countries, as captured by book translations, and two measures of population relatedness. I argue that linguistic distance imposes a cost on idea flows, whereas genetic distance captures an incentive to communicate when dissimilar countries have more to learn from each other. Consistent with this hypothesis, I find that linguistic distance is negatively associated with book translations, whereas genetic distance is positively associated with book translations after conditioning on linguistic and geographic distance. In particular, the benchmark estimate indicates that a one standard deviation increase in linguistic distance reduces book translations by 12 percent, while a one standard deviation increase in genetic distance increases book translations by 10 percent.

Keywords: Linguistic Distance, Genetic Distance, Book Translations, Idea Flows
JEL-Classification: F10, O47, O50, Z10

[†]Brock University, Department of Economics, 1812 Sir Issac Brock Way, L2S 3A2, St. Catharines, ON, Canada (email: adickens@brocku.ca). A special thanks to Nippe Lagerlöf for his detailed feedback, in addition to three anonymous referees. Thanks to Tasso Adamopoulos, Greg Casey, Jeff Chan, Avi Cohen, Oded Galor, Ingo Isphording, and Jeff Quattrociocchi, as well as seminar participants at the Capri Summer School in Economic Growth and York University, for helpful comments. I also thank Ingo Isphording and Sebastian Otten for providing me with the linguistic distance data used in an early draft of this paper, and the Social Science and Humanities Research Council of Canada for financial support.

1 Introduction

Recent research documents a link between the ancestral relationship of two countries and their current difference in income (Spolaore and Wacziarg, 2009). This link is interpreted as reflecting an indirect causal effect: income gaps are smaller between related populations because they are more likely to communicate and adopt similar ideas. By this interpretation the probability of an idea flowing between two countries is the indirect causal link, and this probability is smaller in more distant relationships.

At the same time, dissimilarity could theoretically provide incentive for idea flows if a wider spectrum of non-overlapping traits increase the likelihood of two populations having complementary ideas. This notion of a diversity-driven incentive for idea flows suggests a possible counterbalancing force to the lower probability of communication when two countries are ancestrally distant. This idea is similar in spirit to the theory of Ashraf and Galor (2013), who find that the integration of diverse traits enhances productivity and knowledge creation.[1]

In this paper, I advance the hypothesis that population relatedness confers both social costs and benefits on the cross-country flow of ideas. In particular, I argue that the cost of communication is low between linguistically related populations. Linguistic relatedness thus eases communication and facilitates the flow of ideas. At the same time, the diversity-driven communication incentive is lessened among related populations because shared ancestry implies a similar set of traits and ideas. Genetic relatedness thus reduces the incentive for communication and the spread of new ideas because similar populations have fewer novel ideas to share with each other.[2]

Here, I use book translation data as a measure of international idea flows, and find that relatedness measured across linguistic and genetic dimensions yield robust empirical evidence in support of this hypothesis. In all instances, linguistic distance is negatively associated with book translations. Although the unconditional relationship between genetic distance and book translations is similarly negative, after conditioning on linguistic and geographic distance the association between genetic distance and book translations becomes positive.

I also find that linguistic distance reflects a stronger relationship with book translations than genetic distance in terms of magnitude. Importantly, this difference in magnitude helps reconcile my findings with the argument that related populations are more likely to communicate and adopt similar ideas (Spolaore and Wacziarg, 2009), and the empirical evidence that genetic distance to the technological frontier reflects a barrier to the long-run diffusion

[1] Also see follow-up work by Ashraf et al. (2015) and Ashraf and Galor (2018).

[2] Evidence of a diversity-driven communication incentive is also consistent with the idea that innovation is borne out of isolation (Ashraf et al., 2010), where genetic dissimilarity implies a long history of isolation and consequently a larger set of non-overlapping ideas.

of ideas (Spolaore and Wacziarg, 2014b).[3] As a summary measure of population relatedness, genetic distance captures many different intergenerationally transmitted traits—notably language (Spolaore and Wacziarg, 2016a; Harutyunyan and Özak, 2016). Moreover, the genetic distance between two populations is, in part, determined by geographic distance (Ashraf and Galor, 2013), which itself is a cost to translation flows (Sin, 2017). Unconditional genetic distance negatively correlates with book translations because the costs associated with linguistic and geographic differences supersede the relatively smaller benefits of population differences. Hence the stable and positive relationship between genetic distance and book translations is observable only after separating out the costs associated with linguistic and geographic distance.

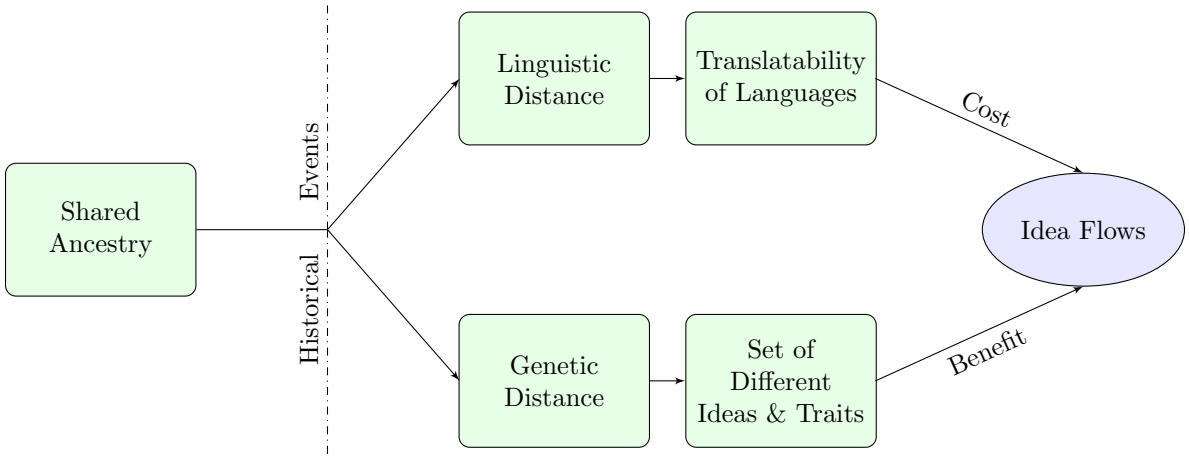
The main data issue to overcome in the analysis is that the origin country of a book translation is not reported in the data. For the baseline estimates, I use the Ethnologue's home country classification for each language, but this means that genetic distance is potentially measured with error if a book translated from an origin language was originally published in a country other than the assigned home country. To be certain this assignment rule is not driving the baseline results, I forego the home-country assumption and instead use a country's share of the global population for each language as a weight to construct a weighted average of every bilateral and country-level covariate used in the baseline, including genetic distance. In effect, these population-weighted covariates reflect a synthetic "language country," where "language country" characteristics measure the population-weighted characteristics of all countries in which a language is spoken. The main finding that population relatedness confers both social costs and benefits on the flow of book translations is robust to this alternative approach.

Using an empirical gravity model of translation flows, I estimate that a one standard deviation increase in linguistic distance yields 12 percent fewer book translations. This result is highly significant and robust to a range of controls, including per capita income and population, political rights, and numerous covariates measuring bilateral differences in geography and colonial history. For the baseline estimate, I flexibly control for country, time and language fixed effects, and later show that the core linguistic distance result holds even when accounting for unobserved country-pair effects. The stability of the linguistic distance estimate in significance and magnitude shows little evidence of a selection bias driving this baseline result.

Similarly, I estimate a significant and robust relationship between book translations and

[3]In this paper, I abstract from a discussion of the direct and barrier effects of culture. Harutyunyan and Özak (2017) highlight the difficulty of disentangling direct and barrier effects and provide evidence of observational equivalence between the two.

Figure 1: Opposing Forces of Population Relatedness on the Flow of Ideas



genetic distance. Only now the sign is reversed, where a one standard deviation increase in genetic distance yields an 10 percent increase in book translations. This result is again robust to a rich set of controls and a variety of robustness checks.

It is intuitive that the positive relationship is observed through genealogical differences. By construction, genetic distance reflects the time since two populations shared a common ancestor. Although linguistic distance reflects the time since two populations shared a common language, this does not necessarily capture the same historical relationship. For example, the Magyar invaded Hungary in the ninth century, imposing their Uralic language on the conquered. To this day Hungarians exhibit a gene distribution similar to the rest of Central Europe, but continue to speak a Uralic language unlike the Latin-based languages of their neighbors (Cavalli-Sforza, 2000). A similar pattern exists in many regions of the transatlantic slave trade, where colonizers imposed their own language on the colonized with little genetic mixing (Phillipson, 1992). Such historical events create a wedge between the co-evolution of language and genetics. In this sense genetic distance is a better summary measure of the time since two populations separated, and thus the extent of dissimilar ideas, beliefs and cultural norms. Figure 1 depicts this argument schematically.

The principal contribution of this paper is the evidence that population relatedness confers both social costs and benefits on the diffusion of knowledge. In related work, Ashraf and Galor (2013) find that diversity confers both costs and benefits on productivity *within* a country.[4] While I am concerned with idea flows and not productivity, the intuition gleaned from Ashraf and Galor (2013) suggests that the interplay between these opposing forces of relatedness also exists *between* countries. I document empirical evidence of this between-

[4]See Ashraf and Galor (2018) for a survey of related work.

country link, which offers a deeper understanding of Spolaore and Wacziarg's (2009) proposed mechanism: more distant populations share fewer ideas overall, with the caveat that this negative relationship operates across linguistic and geographic lines. After conditioning on these costs, ideas flow more readily between dissimilar populations.

By explicitly measuring historical population differences, this paper also speaks to a larger literature on the deep determinants of development. Spolaore and Wacziarg (2013) review this literature, and provide evidence that linguistic and genetic differences can account for the decline in fertility in Europe (Spolaore and Wacziarg, 2014a), create barriers to long-run technology diffusion (Spolaore and Wacziarg, 2014b), and the occurrence of war (Spolaore and Wacziarg, 2016b). Related to this is evidence that the historical composition of a population is a better predictor of its current income than the historical legacy of the geographic location (Putterman and Weil, 2010), that patterns of technology adoption dating back to 1000 BCE persist today and that the effects of past technology on current income is stronger when considering the ancestral composition of a population rather than the population's current location (Comin et al., 2010). At the heart of this literature is the idea that history matters, and that the degree of relatedness is the mechanism linking historical and contemporary development.

The use of book translation data as a measure of idea flows also places this research in proximity to work by Abramitzky and Sin (2014), who document the repressive nature of communist institutions on the inflow of Western books in the Soviet Union prior to its collapse. I account for the institutional environment of a country with a measure of political rights, but find that the coefficient estimates of linguistic and genetic distance are unaltered in significance and magnitude conditional on the extent of political rights. Sin (2017) also finds that geographic distance between countries inhibits the flow of book translations. I also document evidence of this channel, but it holds that the degree of linguistic and genetic distance are contributing factors to the bilateral flow of book translations across countries.

The rest of this paper is structured as follows. Section 2 describes the translation data and details the measurement strategy for each distance measure. I outline the econometric model and report the baseline empirical findings in Section 3, and report numerous robustness checks in Section 4. Section 5 is the discussion section of the paper, where I elaborate on the benefits of dissimilarity. Section 6 concludes.

2 Data

This section outlines the data used to construct the main variables of interest. See Appendix B for more detailed variable definitions, summary statistics and data sources.

2.1 Book Translations as Idea Flows

I use a bibliographic database of book translations from around the world as a measure of international idea flows. Book translations are a recognized measure of idea flows (Abramitzky and Sin, 2014; Sin, 2017), and satisfy the necessary properties of an idea because they are non-rival and disembodied.[5] Book translations are also a suitable measure of idea flows since by definition they require a bilateral exchange in terms of language and often location—a feature that lends itself to bilateral comparisons of population relatedness.

One appealing feature of book translations is the breadth of ideas they capture. Technical ideas are not excluded from the book translation data, but unlike patent citations, these data also capture innovative ideas outside the scope of technology. Books convey ideas and tell stories that reflect different sets of values and cultural norms, and often draw analogy as a form of commentary about the state of society. Rodrik (2014) argues ideas are not only relevant as technical innovations, but also shape our preferences and how we think the world works, and at times "can unlock what otherwise might seem like the iron grip of vested interests" (p. 194).

This behavioural influence of an idea resonates with a long history of books and their influence over society. Emperor Qin Shi Huang famously consolidated the political philosophy of the Qin Dynasty in ancient China, in part, with a wave of book burnings to destroy any writing that challenged his own philosophy. The ceremonial practice of book burning in Nazi Germany was an attempt to rid the country of political writings contrary to the agenda of the National Socialist party. Even more recently, Ayatollah Khomeini issued a fatwā demanding Salman Rushdie be put to death for writing *The Satanic Verses* because it was disrespectful to the Muslim faith. The intolerance and feelings of threat that come from ideas in books speaks to their broader influence on society, politics and our understanding of how the world works.[6]

The influence of books is not only apparent in controversy, but also in their capacity to disseminate ideas. Israel (2009) argues that the transatlantic democratic revolutions of the late eighteenth century have an intellectual origin rooted in the ideas of the Enlightenment, which "persuaded much of the reading elite on either side of the Atlantic [...] that a general revolution in the principle and construction of governments is necessary" (p. 39). This wave

[5] A printed copy of a translated text is a rival good because my purchasing of that book inhibits another's purchase of it. But the translation itself is non-rival because my purchase of the book does not diminish the use of that translation for future copies of the book. This also implies a translation is disembodied in the sense that it is not physically contained as a tangible good.

[6] The empirical evidence that television affects voter turnout (Gentzkow, 2006), social capital (Olken, 2009), the status of women (Jensen and Oster, 2009), fertility decisions (La Ferrara et al., 2012) and much more suggests that books can also influence human behaviour.

of democratic revolutions, Israel argues, was propelled by the spread of pamphlets and books articulating these ideas. This example also speaks to the economic importance of books in the long-run. Given the evidence that democracy causes growth (Acemoglu et al., 2018), by extension the intellectual origin of democracy indirectly links the historical role of books to development patterns of today.

Measuring Book Translations

Translation data was collected from the Index Translationum (IT) database, an international bibliographic archive of book translations hosted by the United Nations Educational, Scientific and Cultural Organization (UNESCO). Since 1932, IT has compiled a detailed record of book translations in print form, and more recently developed an online archive containing bibliographic information of translated books in UNESCO Member States since 1979. Legal deposit legislation states that all publications of a book be submitted to the book repository of a country.[7] Records of book translations are then submitted to IT by the national book repository. The systematic nature of the data collection process is a reassurance of the data's accuracy.

I use data on 1,634,817 book translations that span 119 translating countries over the time period 1979-2005. The panel is unbalanced because I do not observe the same translating country–language pairs in each year. Figure 2 depicts the spatial distribution of observations by quintiles for the baseline sample of translating countries.[8] While the majority of these translations are into the dominant language of the translating country, many countries translate books into more than one language. Each bibliographic entry contains information by subject, the translating country and year, and the origin and target languages of translation.[9] Table B2 lists some of the most translated authors within a random sample of countries, including the subject in which they have most commonly published.

Because the translation data does not report the country in which a book was originally published, I assign a home country to each origin language as stated in the Ethnologue.[10] The benefit of this assignment rule is that it removes judgement and hence any personal bias in the data construction. The drawback is that, in some cases, multiple countries share a common official language so it is not clear the Ethnologue's stated home country

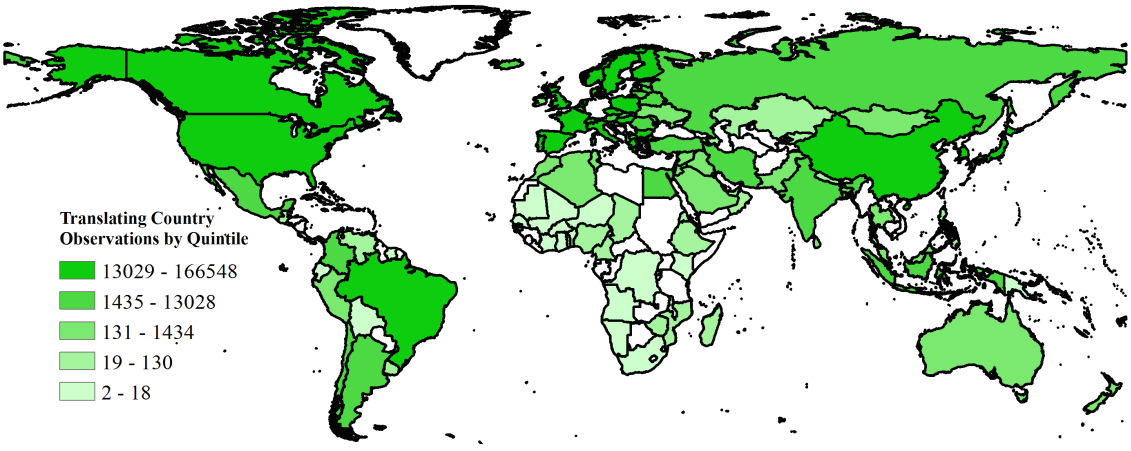
[7] An in-depth report on legal deposit legislation was drafted by UNESCO in 2000, which provides a guideline of how this legislation is formed. See http://archive.ifla.org/VII/s1/gnl/legaldep1.htm.

[8] See Table B3 in the appendix for a complete list of observations by translating country.

[9] Subjects are classified according to the Universal Decimal Classification system, including [1] history, geography and biography, [2] law, social sciences, and education, [3] literature, [4] philosophy and psychology, [5] religion and theology, [6] natural and exact sciences, [7] applied sciences, [8] and arts, games, and sports.

[10] The Ethnologue is a comprehensive database cataloguing all of the world's 7,097 known living languages. It is regarded as the most comprehensive source of its kind. See http://www.ethnologue.com.

Figure 2: Baseline Sample Observations by Country (1979-2005)



is always the correct one. For example, it is unlikely that all translations originating in English were originally published in the United Kingdom as the Ethnologue assignment rule would suggest. In section 4.1 I perform two tests of this assignment rule: I drop the most problematic languages in terms of home country assignment and alternatively construct "language countries" as an average of country characteristics weighted by population shares for each country in which a language is found. In either case the baseline result remains unaltered.

2.2 Language Distance

I use a computerized lexicostatistical measure of linguistic distance that comes from the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology. The starting point of this measure is a set of basic words common across all world languages. For any given language, each word is transcribed according to its pronunciation using a standardized orthography called ASJP-code. Thus far the ASJP team has transcribed this list of words for more than 60 percent of the world's languages. For any two languages of interest, I run a Levenshtein distance algorithm for each word, which in its simplest form calculates the minimum number of edits necessary to translate the spelling of a word from one language to another. A lexicostatistical measure of distance between two languages is calculated as a normalized average of these Levenshtein distances. See Dickens (2018) for a more detailed discussion of the computerized lexicostatistical measure and Appendix B for a formal definition of the measure.

2.3 Genetic Distance

Data on cross-country genetic distance comes from Spolaore and Wacziarg (2009), who collected their data from Cavalli-Sforza et al. (1994). To quantify genetic distance, Cavalli-Sforza et al. (1994) collected data on population allele frequencies specific to a set of selectively neutral genes. An allele is one of many different forms the same gene can assume, where different phenotypic traits (observable characteristics) develop out of different allele sets. Genes were chosen that are known to be selectively neutral to ensure genetic variation across populations is the result of genetic drift. The random nature of drift makes genetic differences simply a function of time. Comparing the distribution of neutral allele frequencies effectively measures the time since two populations shared a common ancestor, so genetic distance becomes a molecular clock.[11] See Appendix B for further discussion of these data.

2.4 Geographic Distance

I also include goegraphic distance in all regressions, measured as the geodesic distance between the most populated cities in a country pair. These data come from the Centre d'Etudes Prospectives et d'Informations Internationales (CEPII). The inclusion of geographic distance accounts for the negative effect of physical distance on translation flows (Sin, 2017), despite the zero trade cost associated with book translations. The suggestion that physical distance is an important determinant of translations flows is consistent with the evidence that the effect of "gravity" is present in the consumption of digital goods over the Internet, which similarly face zero trade costs (Blum and Goldfarb, 2006). Ideas also flow more readily to nearby countries because travel between countries is increasing in proximity, and these cross-border flows of people help facilitate the international diffusion of ideas (Andersen and Dalgaard, 2011).

3 Baseline Estimates

3.1 Econometric Model

Given the bilateral nature of linguistic and genetic distance, I adopt an empirical model similar to the gravity equation. The basic theoretical gravity model implies that bilateral trade between two countries is a function of their economic size and a variety of costs to trade, notably geographic distance. With this in mind, I develop a similar estimation strategy

[11] Because genes are selectively neutral they are independent of natural selection, implying that all conclusions of this paper do not speak to a hierarchy of genetic traits and should not be interpreted this way.

that tests how linguistic and genetic distance affect bilateral translation flows. This basic relationship can be written as:

$$\begin{aligned} TRANS_{ijlt} &= \alpha_0 + \alpha_1 LINGDIST_{jl} + \alpha_2 GENDIST_{ij} + \alpha_3 GEODIST_{ij} \\ &+ \mathbf{X}'_{ij}\mathbf{\Gamma} + \mathbf{X}'_{it}\mathbf{\Omega} + \mathbf{X}'_{jt}\mathbf{\Phi} + \gamma_i + \gamma_j + \gamma_l + \gamma_t + \varepsilon_{ijlt} \end{aligned} \quad (1)$$

where i indexes the translating country, j the origin country (and language), l the target language and t the time period.[12] The dependent variable $TRANS$ measures log translations, $LINGDIST$ and $GENDIST$ measure linguistic and genetic distance, and $GEODIST$ measures the geographic distance in 10,000 kilometre units. $\mathbf{X}_{ij}$ is a vector of bilateral time-invariant measures of geography and colonial relations, while $\mathbf{X}_{it}$ and $\mathbf{X}_{jt}$ include time-varying measures of income, population and political rights in country i and j. A set of fixed effects is also included in each regression, capturing unobserved country effects, time effects and idiosyncratic target language effects. In all regressions I estimate robust standard errors clustered at the country-pair level.

3.2 Unconditional Estimates

In this section I investigate the basic empirical relationship between book translations and three measures of bilateral distance. In particular I investigate how linguistic and genetic distance correlate with book translations in various combinations and disentangle the role of geographic distance.

Table 1 provides summary statistics for these measures. As expected, all three distance measures are positively correlated with each other. Linguistic distance exhibits a positive correlation of 0.34 with genetic distance and 0.24 with geographic distance, and genetic and geographic distance exhibit a positive correlation of 0.45. Book translations exhibit negative correlation with each distance measure, including -0.11, -0.12 and -0.12 with linguistic, genetic and geographic distance. No pairwise correlation between two distance measures is very large in magnitude, and yet all exhibit pairwise negative correlation with book translations, suggesting that each distance measure may have a significant and differential effect on book translations.

Table 2 reports the unconditional estimates of the regression analysis. Columns (1) through (3) indicate that all measure of distance negatively correlate with book translations when separately estimated. The systematic negative relationship between each unconditional distance and book translations is intuitively consistent with the interpretation of Spolaore

[12] It is redundant to denote the origin language since by assumption each origin language is matched to an origin country as previously noted.

Table 1: **Summary Statistics for Distance Measures**

	Simple Correlations			
	Log Translations	Linguistic Distance	Genetic Distance	Geographic Distance
Linguistc distance	-0.11	1.00		
Genetic distance	-0.12	0.34	1.00	
Geographic distance	-0.12	0.24	0.45	1.00
	Summary Statistics			
	Mean	Std dev.	Min	Max
Log translations	1.23	1.59	0.00	8.98
Linguistic distance	0.86	0.12	0.18	1.00
Genetic distance	0.04	0.04	0.00	0.29
Geographic distance (10,000 km)	0.38	0.37	0.01	1.96

42,817 observations for all correlations and summary statistics.

and Wacziarg (2009). The estimate for linguistic distance is not only statistically significant but also economically meaningful: a standard deviation increase in linguistic distance implies an 18.6 percent decrease in book translations. The magnitude of this effect is greater than that of genetic distance, where the estimate in column (2) implies a standard deviation increase in genetic distance yields a 11.9 percent decrease in book translations.

Column (4) reports the estimates from a horse race between all three distance measures. The coefficient estimate for linguistic distance is remarkably stable in magnitude and significance. This is reassuring that linguistic distance is not a latent measure of genetic and geographic differences, but instead an accurate measure of summary language differences. Conversely, after conditioning on both linguistic and geographic distance the correlative relationship between genetic distance and book translations becomes positive. Yet despite this change in sign the estimate remains statistically significant at the 1 percent level. The sensitivity of the coefficient is also suggestive that genetic distance does in fact capture summary population differences, including linguistic and geographic differences.

The influence of geographic distance is also of a notable magnitude: Norway is expected to translate over 3 percent more books originating in neighboring Sweden than Finland

Table 2: Unconditional Baseline Estimates

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-1.72***			-1.52***
	(0.26)			(0.26)
Genetic distance		-3.17***		2.87***
		(1.05)		(0.96)
Geographic distance			-0.86***	-0.85***
			(0.14)	(0.15)
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	42817	42817	42817	42817
Adjusted R^2	0.27	0.26	0.27	0.28
Country pair clusters	2112	2112	2112	2112

This table establishes the unconditional baseline result. Country-pair clustered robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

based on geographic distance alone, all else being equal. On a global scale the influence of geographic distance can become quite large (Sin, 2017).

Taken together, the unconditional estimates of Table 2 nicely summarize the earlier discussion of the opposing forces of relatedness on idea flows. Related populations tend to communicate more ideas because the cost of translation is low, yet dissimilar populations have an incentive to communicate new ideas because the likelihood of complimentary ideas is high.

3.3 Conditional Estimates

To be sure the results of the previous section are not confounded by an omitted variable bias, I test the robustness of the distance measures to a variety of covariates. Column (1) of Table 3 reproduces column (4) of Table 2 using the baseline sample.[13] In column (2) I report estimates that include measures of log real GDP per capita and log population for both the translating and origin country. Adding these time-varying measures leaves the estimates for linguistic and genetic distance almost unchanged in magnitude and statistical significance.

[13]The baseline sample differs because some covariates are not available for every country in each year.

Table 3: Conditional Baseline Estimates

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.55***	-1.55***	-1.54***	-1.27***	-1.24***	-1.13***
	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)
Genetic distance	3.21***	3.19***	3.19***	3.32***	3.28***	2.40**
	(1.04)	(1.04)	(1.05)	(1.03)	(1.02)	(1.05)
Geographic distance	-0.85***	-0.85***	-0.85***	-0.57***	-0.55***	-1.37***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.17)	(0.46)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39275	39275	39275	39275	39275	39275
Adjusted R^2	0.28	0.28	0.28	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897	1897	1897

This table establishes the baseline result for linguistic and genetic distance. Country-pair clustered robust standard errors in parentheses. Economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries, trade controls include the logged value of bilateral trade flows of the country pair, the colonial controls indicate if a country pair has every been in a colonial relationship, and the geography controls indicate contiguity, and a country pair's absolute difference in latitude and longitude. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Similarly, adding political rights has no observable effect on the magnitude and significance of both distances, as shown in column (3). Given that these covariates tend to change at a slow pace this result isn't surprising because country fixed effects absorb the majority of the observable variation.

In column (4) I report estimates that include a measure of log bilateral trade. Bilateral trade flows not only measure the extent of an economic partnership, but they also capture latent determinants of trade such as existing communication networks or the extent of bilateral trust between countries (Guiso et al., 2009). In this sense it is intuitive that a country pair that trades more would also be more inclined to share ideas. Indeed, the coefficient estimate for bilateral trade implies that a 10 percent increase in bilateral trade yields a 1.2 percentage increase in book translations.[14] This suggests that the extent of a bilateral trade relationship has a small positive but significant influence on the level of book translations for a given country pair.

In column (5) I report estimates that include past and present colonial relationships between country pairs, but this doesn't alter the influence of linguistic or genetic distance in any substantial way.

The estimates in column (6) include three measures of geography: an indicator for contiguous country pairs, and a country pair's absolute distance in latitude and longitude. Again, the results are robust to the inclusion of these variables. The positive and significant coefficient on absolute difference in longitude is interesting because it suggests ideas tend to flow between geographically distant countries across similar latitudes. The presence of a north-south friction is consistent with the historical evidence that information tends to flow across east-west axes (Diamond, 1997; Blouin, 2014).[15]

Overall, the opposing forces of linguistic and genetic distance hold even when conditioning on this large set of covariates. The estimates reported in column (6) imply that a standard deviation increase in linguistic distance reduces book translations by 11.7 percent, while a standard deviation increase in genetic distance increases book translations by 10.1 percent. Hereafter I refer to these estimates as my baseline, and the included set of covariates as my baseline set of control variables.

[14] See Table A12 for the bilateral trade coefficient estimate.

[15] Table A12 in the Appendix includes the covariate estimates.

4 Robustness Checks

4.1 Synthetic Language-Country Assignment

A major concern with the baseline estimates is that the home country of a book translation is not known. To overcome this missing information problem, I assign a home country to each origin language as stated in the Ethnologue. However, this assignment rule does not allow for the origin language of a book translation to be assigned to more than one home country, even though books written in the same language may be published in multiple countries. This implies that genetic distance is possibly measured with error. Furthermore, the observed country of where a translation takes place is problematic if publishing houses are geographically concentrated, or if an author of one country decides to publish a book through a publishing house in another country. This again implies genetic distance is possibly measured with error when using the Ethnologue assignment rule.

To address this concern, I construct a synthetic "language country" for every origin and target language of translation that is independent of country borders and instead reflects the entire speaking population of a particular language. I use a country's share of the global population for each language as a weight to construct a weighted average of every bilateral and country-level covariate. This approach is similar in spirit to using Putterman and Weil's (2010) World Migration Matrix to construct ancestry-adjusted country-level data. The resulting unit of observation is a bilateral language pair, where the dependent variable reflects the (logged) global number of translations in a given year for each language pair, and all countries in which each language is spoken are proportionally present in covariate calculations.[16]

Genetic and geographic distance become a population-weighted average measure of distance between every country pairing where the origin and target language are spoken. Bilateral trade and the absolute difference in latitude and longitude are constructed in the same way. The weighted average of a bilateral dummy variable (e.g., if two countries were ever in a colonial relationship) becomes a percentage measure of the indicator for all possible country pairs, and country-level measures such as real GDP and political rights become simple weighted averages. Given the alternative structure of these data, I replace origin and translating country fixed effects in equation (1) with origin and target language fixed effects, and cluster standard errors by language pairs instead of country pairs.

[16]This alternative assignment rule comes at a cost: synthetic language-country calculations require available data for every country in which the translating and origin language is spoken. Consequently, I cannot construct a weighted average with global coverage for every language group because I lack country-level data for a number of countries.

Table 4: Baseline Regressions with Synthetic Language-Country Assignment

	Dependent variable: Log translations				
	(1)	(2)	(3)	(4)	(5)
Linguistic distance	-3.17***			-2.33***	-2.24***
	(0.23)			(0.23)	(0.24)
Genetic distance		-12.99***		5.19***	3.52**
		(1.38)		(1.32)	(1.55)
Geographic distance			-2.64***	-2.61***	-3.04***
			(0.13)	(0.16)	(0.46)
Baseline controls	No	No	No	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	43488	43488	43488	43488	33167
Adjusted R^2	0.53	0.51	0.56	0.58	0.64
Country pair clusters	3249	3249	3249	3249	2690

This table reproduces the baseline estimates using a synthetic language-country assignment scheme for both the home and target language associated with a translation. Home language-country covariates are constructed as a weighted average using a country's share of the global population for a home language as a weight. Translating language-country covariates are similarly constructed using a country's share of the global population for a translating language as a weight. The baseline set of controls included are described in Table 3. Robust standard errors are clustered by language pairs and reported in parentheses.

Table 4 reports the unconditional and conditional population-weighted estimates. The pattern that unfolds is identical to that of Table 2 and 3: the unconditional relationship between each distance measure and book translations is always negative, and genetic distance becomes positive after conditioning on linguistic and geographic distance. The fact that the opposing forces of linguistic and genetic distance are borne out of this alternative assignment rule provides reassurance that the Ethnologue assignment rule is a good approximation of a book translation's country of origin. In fact, the estimates reported in Table 4 are larger in magnitude than the baseline estimates, suggesting that the Ethnologue assignment rule is the conservative approach to take with respect to the missing information problem.

Problematic Languages

A related concern of the Ethnologue assignment rule is that some languages are so widely spoken that it's difficult to pin them down to a single country. Here, I identify the most

problematic languages and purge them from the data to test the robustness of the baseline results. First, I use the Ethnologue to generate a list of languages sorted by first-language speakers. The top five languages by number of speakers include Mandarin (1,197 million), Spanish (414 million), English (335 million), Hindi (260 million) and Arabic (237 million). Second, I use the Ethnologue to generate a list of languages sorted by the number of countries in which the language is spoken. The top five languages by this definition include English (99 countries), Arabic (60 countries), French (51 countries), Mandarin (33 countries) and Spanish (31 countries). The result is a list of six potentially problematic languages: English, Arabic, French, Mandarin, Spanish and Hindi.

I proceed by systematically dropping all book translations to or from a problematic language, using the baseline dataset (Panel A) and the synthetic language-country dataset (Panel B), and report these results in Table 7. All estimates are statistically significant with the expected sign, and the majority are comparable to the baseline estimate in terms of magnitude. It is important to note that the estimates are robust to excluding all book translations to or from English, the most problematic language to assign to a single country. Even when excluding all six problematic languages the estimates remain robust—a sample restriction that amounts to more than a 30 percent drop in observations relative to the baseline sample. Taken together, the results in Table 7 suggest that the Ethnologue country assignment rule is not driving the baseline results.

4.2 Within-Continent Correlation

An additional concern of the empirical strategy used here is the possible strong association between genetic and geographic distance within continents. Serial founder effects imply that the genetic distance between two populations is increasing in geographic distance along the historical migratory route out of East Africa. The implication being that, within continents, the distance from the out-of-Africa migratory route is positively correlated with the distance between countries, and thus correlated with genetic distance.

Table 5 reports within-continent correlations for all key distance variables. The pairwise correlation coefficient for genetic and geographic distance is now 0.56, relative to the unconditional correlation coefficient of 0.45 reported in Table 1. While the increase is expected, the modest change is likely due to the fact that the data include country pairs from around the world (i.e., within-continent and across-continent bilateral pairs). Furthermore, post-1500 population movements such as the slave trade and colonization of the New World have worked to partially break the strong association between geographic and genetic distance.

Nonetheless, to be certain that the genetic distance result is not a consequence of this

Table 5: Robustness Check for Problematic Languages of Translation

Excluding:	Dependent variable: Log translations						
	English Language Translations	Arabic Language Translations	French Language Translations	Mandarin Language Translations	Spanish Language Translations	Hindi Language Translations	All Six Language Translations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Ethnologue Country Assignment						
Linguistic distance	-1.89***	-1.07***	-1.07***	-1.13***	-1.09***	-1.09***	-1.69***
	(0.27)	(0.29)	(0.28)	(0.28)	(0.28)	(0.29)	(0.26)
Genetic distance	3.17***	2.43**	1.88*	2.65**	1.88*	2.35**	2.33**
	(0.99)	(1.04)	(1.02)	(1.07)	(1.10)	(1.04)	(0.94)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	28742	37062	33138	37885	35429	38791	18178
Adjusted R^2	0.38	0.29	0.31	0.28	0.29	0.28	0.44
Country pair clusters	1677	1725	1796	1868	1745	1891	1235
	Panel B: Synthetic Language-Country Assignment						
Linguistic distance	-1.99***	-2.28***	-2.17***	-2.24***	-2.10***	-2.26***	-1.79***
	(0.24)	(0.24)	(0.25)	(0.24)	(0.25)	(0.25)	(0.25)
Genetic distance	3.28**	3.62**	3.26**	3.79**	3.96***	3.43**	3.98***
	(1.46)	(1.57)	(1.51)	(1.64)	(1.38)	(1.54)	(1.30)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	29111	31830	30083	31903	31134	32236	21097
Adjusted R^2	0.61	0.64	0.61	0.64	0.63	0.64	0.59
Country pair clusters	2449	2607	2503	2609	2593	2624	1955

This table establishes that the baseline results are robust to excluding commonly spoken and multinational languages from the data. Panel A reports estimates for the baseline sample, where standard errors are clustered by country pairs. Panel B reports estimates for the synthetic language-country sample, where standard errors are clustered by language pairs. All reported estimates also include controls for geographic distance and the baseline set of control variables used in Table 3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Within-Continent Correlation

	Log Translations	Linguistic Distance	Genetic Distance	Geographic Distance
Linguistc distance	-0.10	1.00		
Genetic distance	-0.10	0.31	1.00	
Geographic distance	-0.09	0.29	0.56	1.00

42,817 observations for within-continent correlations.

within-continent correlation, I replicate my baseline estimates with and without geographic distance. Columns (1) and (2) in Table 6 report these results. The coefficients on genetic distance vary little across specification, suggesting that the correlation with geographic distance cannot explain away these results.[17]

I also test for heterogeneity with respect to country pairs that are both ancestral to Europe, given that the east-west spread of countries across the continent correlates with the historical east-west migratory route across the continent. I use Putterman and Weil's (2010) World Migration Matrix to identify country pairs where 75 percent or more of the population in both countries are ancestral to Europe. Columns (3) and (4) of Table 6 report these results. The fact that the positive effect of genetic distance only holds among country pairs ancestral to Europe suggests that controlling for geographic distance may be problematic. However, the results in columns (4) suggest that the positive and significant coefficient on genetic distance is quite stable even in the absence of geographic distance. Taken together, the estimates suggest that the diversity-driven communication incentive exists between populations with large European ancestry, but that this result is not an outcome of the correlation between genetic and geographic distance.

4.3 Additional Robustness Checks

In Appendix A, I also show that the results are robust to a variety of specifications. I test the home country assumption further by systematically dropping every origin language of a translation with at least 100 observations in the benchmark sample (Table A1). I find no evidence of non-linearities in linguistic and genetic distance (Table A2), nor any evidence of structural language differences driving the results (Table A3). I disaggregate

[17] Comparing the coefficients for genetic distance across specifications (1) and (2) yields a $\chi^2 = 0.92$, implying that I cannot reject the null hypothesis of statistical equivalence.

Table 7: Robustness Check for Within-Continent Correlation

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-1.13***	-1.12***	-1.02***	-1.01***
	(0.28)	(0.28)	(0.36)	(0.37)
Genetic distance	2.40**	2.22**	-0.14	-0.06
	(1.05)	(1.03)	(1.44)	(1.31)
Geographic distance	-1.37***		-1.30***	
	(0.46)		(0.45)	
Linguistic distance			-0.18	-0.17
× European ancestry			(0.48)	(0.48)
Genetic distance			3.37**	3.33**
× European ancestry			(1.63)	(1.57)
Geographic distance			-0.22	
× European ancestry			(0.17)	
European ancestry			-0.10	-0.23
			(0.43)	(0.43)
Baseline controls	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	39275	39275	39275	39275
Adjusted R^2	0.28	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897

This table establishes that within-continent correlation is not driving the baseline result. Country-pair clustered robust standard errors in parentheses. All regressions include the baseline set of control variables used in Table 3. * p < 0.10, ** p < 0.05, *** p < 0.01.

book translations by economic and cultural idea types, and although the main result holds, the coefficient estimates are statistically equivalent across idea types (Table A4). I also account for differences in human capital across countries (Table A5) and existing bilateral relationships between country pairs (Table A6). I collapse the data into a cross-section to address the fact that there is less variation across time than across countries (Table A7) and replicate my baseline estimates using an alternative measure of linguistic distance (Table A8). As a final check, I systematically drop translations by book subject and find no evidence that a single subject is driving the results (Table A9).

5 Discussion

5.1 On the Benefits of Dissimilarity

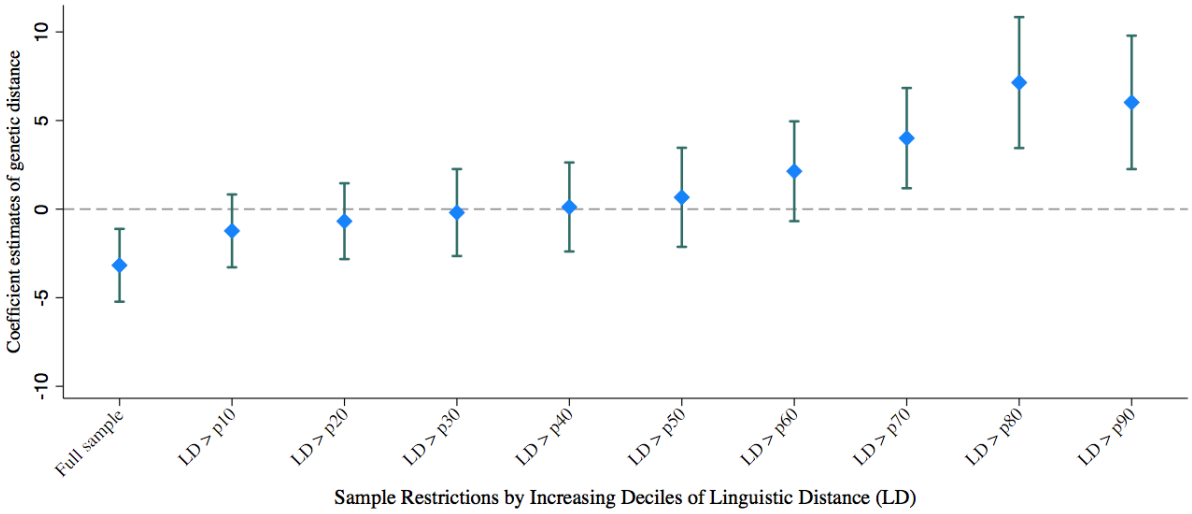
At what level of linguistic distance does genetic distance become positive? To shed light on this question, I explore the unconditional relationship between book translations and genetic distance using subsamples of the data, where the included country-language pairs in each subsample are restricted by increasing thresholds of linguistic distance. For this exercise, I estimate a simplified version of equation (1) that excludes all other control variables, including linguistic and geographic distance. Figure 3 depicts these subsample estimates.[18]

For the full sample estimate, the coefficient on genetic distance is negative and statistically significant—this estimate is equivalent to the reported estimate in column (2) of Table 2. For the subsequence estimates, I divide the observed distribution of linguistic distance into deciles, and systematically drop observations below each decile. Moving from left to right along the x-axis of Figure 3, the estimates come from increasingly smaller subsamples that are composed of more linguistically distant country-language pairs. A near-monotonic relationship unfolds, where genetic dissimilarity becomes beneficial to translation flows for observations in the top quartile of the distribution of linguistic distance.

Table A10 in the Appendix reports estimates of the full regression model including an interaction term between linguistic and genetic distance. The estimates in column (2) imply that the marginal effect of genetic distance becomes positive at a threshold of 87 percent dissimilarity in language, a level just past the sample mean of 86 percent. As a further check, I separate the sample into two subsamples: observations above and below the 87 percent threshold. Column (3) reports estimates for observations above this threshold, where the estimate for genetic distance is positive and significant. To the contrary, the coefficient on genetic distance is negative but statistically no different than zero for the sample of

[18] Table A11 in the Appendix reports the regression coefficients.

Figure 3: Estimates of Unconditional Genetic Distance



This figure depicts unconditional estimates of genetic distance for 10 different samples. The sample is cut by deciles of linguistic distance, and moving from left to right, the estimates of genetic distance come from increasingly smaller subsamples that are made up of more linguistically distant country-language pairs. Intervals reflect 95% confidence levels.

observations below this threshold, as shown in column (4). Taken together, these results confirm that the benefits of genetic distance are conditional on the degree of cross-country language differences.

6 Concluding Remarks

This paper is motivated by the observation that linguistic and genetic distance affect cross-country income differences (Spolaore and Wacziarg, 2009). Income gaps are thought to be smaller between related populations because they are more likely to communicate and adopt similar ideas than dissimilar populations. I test this interpretation using book translation data, and find that more distant populations do indeed share fewer ideas overall, with the caveat that this negative relationship operates across linguistic and geographic lines. Yet conditional on these linguistic and geographic differences, ideas flow more readily between dissimilar populations. This speaks to the evidence of two opposing forces of relatedness found to exist *within* countries (Ashraf and Galor, 2013), and contributes to this literature with evidence that these opposing forces of relatedness also exist *between* countries.

To reconcile this evidence with the interpretation of Spolaore and Wacziarg (2009), I also show that linguistic and geographic distance reflect a stronger relationship with book

translations than genetic distance in terms of magnitude. Unconditional genetic distance negatively correlates with book translations because the costs associated with linguistic and geographic differences supersede the relatively smaller benefits of population differences. The stable and positive relationship between genetic distance and book translations is observable only after separating out the costs associated with linguistic and geographic distance.

Overall, these findings document the important role of population relatedness in the diffusion of knowledge. Recognizing that the empirical evidence here speaks to book translations in only the last few decades, I believe the core result is informative of a more general relationship between idea flows and population relatedness. This empirical finding is important because it suggests one society's exposure and interaction with new ideas is not only determined by the pool of other countries that share similar histories, but also the type of shared history (e.g., linguistic or biological). The importance of distinguishing between the type of shared history is evident in the opposing relationship linguistic and genetic distance exhibit with book translations. Hence, the benefits of a more integrated global network of idea sharing may be achieved by overcoming linguistic barriers with directed education policy, and by improving incentives for the translation of ideas across borders.

References

Abramitzky, R. and Sin, I. (2014). Book Translations As Idea Flows: The Effects of the Collapse of Communism on the Diffusion of Knowledge. *Journal of the European Economic Association*, 12(6):1453–1520.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2018). Democracy Does Cause Growth. *Journal of Political Economy*, Forthcoming.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, 8(2):155–194.

Andersen, T. B. and Dalgaard, C.-J. (2011). Flows of People, Flows of Ideas, and the Inequality of Nations. *Journal of Economic Growth*, 16(1):1–32.

Ashraf, Q. and Galor, O. (2013). The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development. *American Economic Review*, 103(1):1–46.

Ashraf, Q. and Galor, O. (2018). The Macrogenoeconomics of Comparative Development. *Journal of Economic Literature*, Forthcoming.

Ashraf, Q., Galor, O., and Klemp, M. (2015). Heterogeneity and Productivity. *Brown University, mimeo.*

Ashraf, Q., Galor, O., and Özak, Ö. (2010). Isolation and development. *Journal of the European Economic Association*, 8(2/3):401–412.

Bakker, D., Brown, C. H., Brown, P., Egorov, D., Grant, A., Holman, E. W., Mailhammer, R., Müller, A., Velupillai, V., and Wichmann, S. (2009). Add Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology*, 13:167–179.

Barbieri, K., Keshk, O. M. G., and Pollin, B. (2009). Trading Data: Evaluating our Assumptions and Coding Rules. *Conflict Management and Peace Science*, 26(5):471–491.

Barro, R. and Lee, J.-W. (2013). A New Data Set of Education Attainment in the World, 1950-2010. *Journal of Development Economics*, 104(1):184–198.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Blouin, A. (2014). Culture, Isolation, and the Diffusion of Knowledge: Evidence from the Bantu Expansion. *University of Toronto, mimeo.*

Blum, B. S. and Goldfarb, A. (2006). Does the Internet Defy the Law of Gravity? *Journal of International Economics*, 70(2):384–405.

Cavalli-Sforza, L. (2000). *Genes, Peoples, and Languages*. North Point Press, New York.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton.

Comin, D., Easterly, W., and Gong, E. (2010). Was the Wealth of Nations Determined in 1000 BC? *American Economic Journal: Macroeconomics*, 2(3):65–97.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2012). The Political Economy of Linguistic Cleavages. *Journal of Development Economics*, 97(2):322–338.

Diamond, J. (1997). *Guns, Germs and Steel*. W. W. Norton & Company, New York.

Dickens, A. (2018). Ethnolinguistic Favoritism in African Politics. *American Economic Journal: Applied Economics*, 10(3):370–402.

Fearon, J. D. (2003). Ethnic and Cultural Diversity by Country. *Journal of Economic Growth*, 8(2):195–222.

Gentzkow, M. (2006). Television and Voter Turnout. *Quarterly Journal of Economics*, 121(3):931–971.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural Biases in Economic Exchange? *The Quarterly Journal of Economics*, 124(3):1095–1131.

Harutyunyan, A. and Özak, Ö. (2016). Culture, Diffusion and Economic Development. *SMU mimeo*.

Harutyunyan, A. and Özak, Ö. (2017). Culture, Diffusion and Economic Development: The Problem of Observational Equivalence. *Economics Letters*, 158:94–100.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2009). Explorations in Automated Language Classification. *Folia Linguistica*, 42(3-4):331–354.

Israel, J. (2009). *A Revolution of the Mind: Radical Enlightenment and the Intellectual Origins of Modern Democracy*. Princeton University Press, Princeton.

Jensen, R. and Oster, E. (2009). The Power of TV: Cable Television and Women's Status in India. *The Quarterly Journal of Economics*, 124(3):1057–1094.

La Ferrara, E., Chong, A., and Duryea, S. (2012). Soap Operas and Fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, 4(4):1–31.

Olken, B. A. (2009). Do Television and Radio Destroy Social Capital? *American Economic Journal: Applied Economics*, 1(4):1–33.

Phillipson, R. (1992). *Linguistic Imperialism*. Oxford University Press, Oxford.

Psacharopoulos, G. (1994). Returns to Investment in Education: A Global Update. *World Development*, 22(9):1325–1343.

Putterman, L. and Weil, D. N. (2010). Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality. *The Quarterly Journal of Economics*, 125(4):1627–1682.

Rodrik, D. (2014). When Ideas Trump Interests: Preferences, Worldviews, and Policy Innovations. *Journal of Economic Perspectives*, 28(1):189–208.

Sin, I. (2017). The Gravity of Ideas: How distance affects translations. *The Economic Journal*.

Spolaore, E. and Wacziarg, R. (2009). The Diffusion of Development. *The Quarterly Journal of Economics*, 124(2):469–529.

Spolaore, E. and Wacziarg, R. (2013). How Deep Are the Roots of Economic Development? *Journal of Economic Literature*, 51(2):325–369.

Spolaore, E. and Wacziarg, R. (2014a). Fertility and Modernity. *Tufts University Discussion Papers Series 0779*.

Spolaore, E. and Wacziarg, R. (2014b). Long-Term Barriers to Economic Development. *Handbook of Economic Growth*, 2:121–176.

Spolaore, E. and Wacziarg, R. (2016a). Ancestry, Language and Culture. In Ginsburgh, V. and Weber, S., editors, *The Palgrave Handbook of Economics and Language*, pages 174–211. Palgrave Macmillan, London.

Spolaore, E. and Wacziarg, R. (2016b). War and Relatedness. *Review of Economics and Statistics*, 98(5):925–939.

Swadesh, M. (1952). Lexicostatistical Dating of Prehistoric Ethnic Contracts. *Proceedings of the American Philosophical Society*, 96:121–137.

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21:121–137.

Wichmann, S., Holman, E. W., Bakker, D., and Brown, C. H. (2010). Evaluating Linguistic Distance Measures. *Physica A*, 389(17):3632–3639.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., and Valenzuela, P. (2013). The ASJP Database. Version 16.

FOR ONLINE PUBLICATION

A Supplementary Tables

A.1 Two Further Tests of the Home Country Assumption

In this section I extend the analysis of section 4.1 and systematically dropping every origin language of a translation with at least 100 observations in the benchmark sample. The idea here is to ensure that the benchmark results cannot be explained away by any one origin language due to the ambiguity of the home country assignment rule. Table A1 reports these estimates and shows no evidence that any one language can explain away the benchmark result.

A.2 Non-Linear Relationships

As an additional robustness check, I test for the possibility of non-linearities in each of the three distance measures (Ashraf and Galor, 2013). For this test I augment the baseline model to include a quadratic term. Table A2 reports these results. In columns (1) through (3) I include the quadratic for each distance measure individually, and in column (4) I include all three quadratics. I find no evidence of a non-linear relationship for any of the distance measures.

A.3 Differences in Across Country Language Structure

One additional concern is that countries are structurally different in terms of language. For example, multilingual countries may translate more books because they are more linguistically diverse. Similarly, the colonial legacy of language is evident in the fact that many countries still use their colonizer's native language as a regional lingua franca. This may influence both the number of books translated and specific languages that are commonly translated. A shared language between two countries may also influence how much those two countries translate each other's writings. In this section, I investigate further sample restrictions to be sure the benchmark results are not driven by country-level differences in language structure. Results are reported in Table A3, where all estimates include a full set of controls.

Column (1) reports model estimates from a subsample of unilingual countries. The estimated coefficient for linguistic and genetic distance are consistent with the previous estimates in terms of significance and magnitude.

Next I investigate the influence of lingua francas within a translating country. Using data from Alesina et al. (2003), I identify all target languages that are considered to be a lingua franca in the *translating* country and exclude those observations from the regression. The virtue of this approach is that a language will not always be considered a lingua franca in every country, only in cases when it is commonly used by a significant portion of the population as a bridging language. Column (2) reports these estimates. Once again, after dropping all books translated into a lingua franca of the translating country, model estimates are qualitatively and quantitatively consistent with the benchmark estimate.

In column (3) I drop all observations for country pairs that share a common official language, and in column (4) I drop all observations with at least 9 percent of the population in both countries speaking the same language. Again the benchmark result holds.

Column (5) is the most restrictive test, where I drop all translations including multilingual countries, lingua franca translations and shared language countries (by either definition). The results conform with the benchmark, albeit with a slightly larger magnitude of influence in terms of linguistic distance. Despite this difference the same qualitative result holds.

A A.4 Distance Effects by Idea Types

Ashraf and Galor (2013) find that diverse societies produce more scientific ideas and are more productive than homogeneous societies as a result. The wider spectrum of non-overlapping traits that are characteristic of these diverse societies increase the likelihood that different ideas are complementary to the advancement of technology. By this same logic, genetically distant countries have a large incentive to communicate economic ideas because distant ideas are newer and are more likely to be productivity-enhancing.

To explore the possibility that economic ideas are more responsive to genetic distance, I extend the analysis to include two categories of idea type: economic and cultural ideas. Book translations pertaining to the applied and natural sciences are coded as having economic use value, whereas translations pertaining to the social sciences, philosophy, history, literature, religion and the arts are coded as having cultural use value.

To preserve the sample size of the baseline estimates I construct the dependent variable for economic book translations as $\ln(1 + \text{economic translations})$ and the dependent variable for cultural book translations as $\ln(1 + \text{cultural translations})$. In other words, if a country only translates cultural books in a given year, the dependent variable for economic ideas equals zero where it would have otherwise taken a strictly positive value in the baseline model. The opposite is true for the measure of cultural book translations.

Table A4 reports the standardized beta coefficients for each idea type to allow for compa-

rability across coefficients. Consistent with the baseline estimates, both linguistic and genetic distance are statistically significant with the expected sign irrespective of idea type. These results suggest that the opposing forces of relatedness are still at play when disaggregating books by their economic and cultural use value.

The standardized beta coefficient for genetic distance is 0.06 for cultural book translations, compared to economic book translations that yield a standardized beta coefficient of 0.07. While these estimates do corroborate the interpretation and empirical results of Ashraf and Galor (2013), the coefficients are statistically equivalent to one another. The fact that the coefficients are not statistically different may reflect the fact that genetically distant countries have an equally large incentive to communicate cultural ideas as they do economic ideas. Nonetheless, what's important is that these results do not contradict the main result of the paper—that the robust positive association between genetic distance and book translations holds irrespective of idea type.

A.5 Human Capital

A country's decision to translate might also be influenced by its level of human capital, since educated populations are more likely to have a high demand for book translations. Using data from Barro and Lee (2013), I measure human capital in the translating country in four ways: a country's average years of schooling attained, the percentage of the total population who have completed primary schooling, secondary schooling and tertiary schooling. Because these data are only available in 5-year intervals I supplement it with data from the Penn World Tables version 8.0. I use a yearly index of human capital per person, based on years of school (Barro and Lee, 2013) and the returns to education (Psacharopoulos, 1994). Table A5 reports robustness checks using these different measures of human capital.

Column (1) includes the Penn World Table human capital index, which I match to 99 percent of my benchmark sample. The translating country human capital index positively correlates to book translations as expected, but is estimated to be no different than zero. The linguistic and genetic distance estimates are robust to the inclusion of human capital, and quantitatively equivalent to the benchmark estimates. Columns (2) through (6) add each original Barro-Lee measure and again the variables of interest are unaltered. In column (3) the human capital measure is estimated with the wrong sign, but in all other cases the expected positive sign results. Overall these results suggest human capital cannot explain away the benchmark estimates.

A.6 Existing Bilateral Relationships

The most effective way to capture any relevant time-invariant feature of a country pair is to include country pair fixed effects. The major limitation of this approach is that it is no longer possible to explicitly estimate genetic distance because country pair variation is time-invariant and adsorbed by the fixed effects estimator.[19] Nonetheless I proceed to test the robustness of the benchmark linguistic distance estimate.

Table A6 reports the benchmark estimate with country pair fixed effects in place of individual country effects. Column (1) reports estimates where I forego any covariates in order to maximize the sample. Linguistic distance is estimated to be significantly different from zero with the expected sign and with a magnitude of influence similar to the benchmark estimate. I then incrementally add the benchmark covariates that are not absorbed by the country-pair fixed effects. Overall the results are similar to the benchmark, albeit a little noiser and consequently less precisely estimated. However, the robustness of the linguistic distance estimate to country-pair fixed effects confirms that time-invariant country-pair variation cannot explain away the benchmark estimate.

A.7 Check for Understated Standard Errors

Because linguistic and genetic distance are assumed to be constant over the sampled time period, standard errors may be understated due to repeated values of each distance measure for the same language-country pair in the panel (Bertrand et al., 2004). To test for this I regress log translations on a set of year dummies and collapse the data by averaging over time the residual translation variation. Similarly, all other time-variant independent variables are regressed on time dummies and the residuals are collapsed by averaging over time. The benchmark specification is then re-estimated using this collapsed and time-averaged data. Results are reported in table A7. Linguistic and genetic distance are again estimated significantly significant and have the expected signs.

A.8 Alternative Measure of Linguistic Distance

In this section I reproduce the table of benchmark estimates using a cladistic measure of linguistic distance. I construct this measure of cladistic distance as outlined in Appendix B, which is measured as one minus the ratio of shared branches on the Ethnologue language tree.

[19] Although language distances are assumed to be constant over the sampled time period, country-pair fixed effects do not absorb language distance effects because each unit of observation is indexed by country-pair, year *and* the target language of translation – the latter of which isn't constant for a country pair each year.

I assume the weighting scheme of Fearon (2003), where the ratio of shared tree branches is square rooted to discount more recent linguistic cleavages relative to deep cleavages. Because the number of branches varies among language familes and subfamilies, the maximum number of branches between any two languages is not constant. To overcome this obstacle I assume that all current languages are of equal distance from the proto-language at the root of the Ethnologue language tree. This assumption is equivalent to the assumption Desmet et al. (2012) use when constructing cladistic distances (see Figure B1).

Table A8 reports the benchmark estimates using this alternative measure. The two opposing forces of relatedness are borne out of this alternative data and estimated to be significantly different than zero. The effect of the cladistic measure of linguistic distance is smaller than the comparable lexicostatistical measure used in the benchmark estimate. This finding is consistent with the evidence in Dickens (2018), where the added variation of the lexicostatistical measure yields added percision in estimation. Nonetheless, the basic finding of this paper is robust to alternative measures of linguistic distance.

A.9 Dominant Book Subject?

A potential concern is that the interplay between linguistic and genetic distance observed in the benchmark estimate is the result of a strong statistical association with one field of study in the aggregate translation data. Regression analysis using total translations as the dependent variable may hide the fact that certain idea types are strongly influenced by distance while others are not. If so, then excluding book translations from the dependent variable that belong to some outlier subject would yield estimates of linguistic and genetic distance substantially different from the baseline.

To test this I re-estimate the preferred specification and drop all translations belonging to each subject area one at a time. The results are presented in Table A9. Coefficient estimates should be interpreted relative to the benchmark estimate; i.e., small deviations from the benchmark estimate imply the excluded subject of translation is not influential over and above the average effect, whereas large deviations indicate subject areas that are particularly influential in the benchmark result.

The first observation about Table A9 is that no one subject area is driving the core result of this study. Second, both genetic and linguistic distance are precisely estimated in all regressions at standard levels of confidence. These two observations suggest the baseline results of section 3 cannot be explained by a single outlier subject that is particularly responsive to either linguistic or genetic distance.

Table A1: Sensitivity Analysis: Further Test of Home Country Assignment

Language Dropped	Linguistic Distance	Genetic Distance	*N*
German	-1.29***	2.15**	36,498
	(0.28)	(1.04)	
Greek	-1.12***	2.77**	36,920
	(0.28)	(1.10)	
Dutch	-1.03***	2.42**	38,020
	(0.28)	(1.06)	
Swedish	-1.06***	2.36**	38,060
	(0.29)	(1.06)	
Russian	-1.11***	2.27**	38,158
	(0.29)	(1.06)	
Danish	-1.08***	2.44**	38,205
	(0.29)	(1.06)	
Portuguese	-1.12***	2.54**	38,211
	(0.29)	(1.08)	
Japanese	-1.10***	2.08**	38,221
	(0.28)	(1.02)	
Hebrew	-1.14***	2.39**	38,261
	(0.28)	(1.06)	
Polish	-1.15***	2.60**	38,284
	(0.29)	(1.06)	
Norwegian	-1.05***	2.44**	38,373
	(0.29)	(1.06)	
Hungarian	-1.14***	2.40**	38,499
	(0.28)	(1.16)	
Persian (Iranian)	-1.11***	2.41**	38,620
	(0.28)	(1.04)	
Finnish	-1.11***	3.03***	38,634
	(0.28)	(1.07)	
Romanian	-1.17***	2.46**	38,675
	(0.29)	(1.05)	
Turkish	-1.12***	2.29**	38,704
	(0.28)	(1.05)	
Sanskrit	-1.13***	2.40**	38,756
	(0.28)	(1.04)	
Bengali	-1.10***	2.27**	38,782
	(0.29)	(1.05)	
Czech	-1.11***	2.39**	38,785
	(0.29)	(1.05)	
Bulgarian	-1.13***	2.39**	38,876
	(0.29)	(1.05)	
Tibetan	-1.13***	2.27**	38,887
	(0.28)	(1.07)	
Yiddish	-1.12***	2.41**	38,897
	(0.29)	(1.05)	
Icelandic	-1.11***	2.44**	38,907
	(0.29)	(1.05)	
Urdu	-1.11***	2.37**	38,941
	(0.29)	(1.05)	
Catalan	-1.13***	2.43**	38,948
	(0.29)	(1.06)	
Korean	-1.12***	2.63**	38,993
	(0.28)	(1.09)	
Albanian	-1.12***	2.37**	39,011
	(0.29)	(1.05)	
Croatian	-1.13***	2.38**	39,023
	(0.29)	(1.05)	
Vietnamese	-1.13***	2.44**	39,071
	(0.29)	(1.05)	
Slovene	-1.15***	2.41**	39,080
	(0.29)	(1.05)	
Pali	-1.13***	2.39**	39,089
	(0.28)	(1.05)	
Slovak	-1.12***	2.40**	39,099
	(0.29)	(1.05)	
Estonian	-1.12***	2.48**	39,112
	(0.28)	(1.06)	
Indonesian	-1.13***	2.37**	39,115
	(0.28)	(1.05)	
Syriac	-1.12***	2.38**	39,120
	(0.28)	(1.05)	
Panjabi	-1.12***	2.38**	39,136
	(0.28)	(1.05)	
Irish	-1.13***	2.41**	39,137
	(0.28)	(1.05)	
Arabic (Egyptian)	-1.13***	2.40**	39,139
	(0.28)	(1.05)	
Ukrainian	-1.15***	2.45**	39,145
	(0.28)	(1.05)	
Aramaic	-1.12***	2.41**	39,148
	(0.28)	(1.05)	
Afrikaans	-1.11***	2.46**	39,149
	(0.28)	(1.05)	
Lithuanian	-1.12***	2.41**	39,160
	(0.28)	(1.06)	
Coptic	-1.13***	2.40**	39,165
	(0.28)	(1.05)	
Thai	-1.13***	2.35**	39,166
	(0.28)	(1.05)	
Latvian	-1.13***	2.41**	39,171
	(0.28)	(1.05)	

This table tests the assignment of the home country of a language by systematically dropping each original language of a book translation with a 100 or more observations. The robustness of the results to this test suggests the benchmark result is not driven by the assumption of the original country of a translation. All regressions include the benchmark set of control variables used in column (6) of Table 3. All regressions also include individual country, year and target language fixed effects. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Robustness Check for Non-Linearities

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-4.02*	-1.14***	-1.13***	-3.99*
	(2.13)	(0.28)	(0.28)	(2.14)
Genetic distance	2.12**	4.50*	2.45**	4.39*
	(1.05)	(2.40)	(1.06)	(2.36)
Geographic distance	-1.36***	-1.39***	-1.41***	-1.47***
	(0.46)	(0.46)	(0.52)	(0.52)
Linguistic distance squared	1.99			1.96
	(1.50)			(1.51)
Genetic distance squared		-16.94		-17.32
		(17.09)		(16.56)
Geographic distance squared			0.03	0.06
			(0.19)	(0.19)
All controls	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	39275	39275	39275	39275
Adjusted R^2	0.28	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897

Country-pair clustered robust standard errors in parentheses. All regressions include the full set of control variables used in Table 3. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A3: Robustness Check for Differences in Across-Country Language Structure

	Dependent variable: Log translations				
	(1)	(2)	(3)	(4)	(5)
Excluding:	Multilingual Countries	Lingua Franca Translations	Common Language Countries (official)	Common Language Countries (> 9% pop.)	All Excluded Countries
Linguistic distance	-1.18***	-1.10***	-1.43***	-1.41***	-1.37***
	(0.32)	(0.28)	(0.30)	(0.31)	(0.33)
Genetic distance	2.82**	2.43**	2.91***	2.92***	2.63**
	(1.13)	(1.07)	(1.07)	(1.07)	(1.13)
Geographic distance	-1.26***	-1.27***	-1.44***	-1.23***	-1.39***
	(0.49)	(0.45)	(0.42)	(0.42)	(0.46)
Benchmark controls	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	32,588	37,628	34,481	33,589	28,279
Adjusted R^2	0.29	0.29	0.31	0.31	0.31
Country pair clusters	1636	1840	1743	1707	1440

This table reports estimates from various subsamples that exclude potentially problematic countries and translations. All regressions include the benchmark set of control variables used in column (6) of Table 3. Country-pair clustered robust standard errors reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A4: Distance Effects by Cultural and Economic Idea Types

	Cultural book translations		Economic book translations	
	(1)	(2)	(3)	(4)
Linguistic distance	-0.10***	-0.07***	-0.12***	-0.09***
	(0.02)	(0.02)	(0.03)	(0.02)
Genetic distance	0.08***	0.06**	0.09***	0.07**
	(0.03)	(0.03)	(0.03)	(0.03)
Geographic distance	Yes	Yes	Yes	Yes
Baseline controls	No	Yes	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	39275	39275	39275	39275
Adjusted R^2	0.26	0.27	0.25	0.26
Country pair clusters	1897	1897	1897	1897

This table establishes that the two opposing forces of relatedness exist across different idea types. Estimates are reported as standardized beta coefficients to allow for comparability across regressions. Cultural book translations are those from the field of social sciences, philosophy, history, literature, religion and the arts. Economic book translations are those from the fields of natural and applied sciences. Control variables include the full set of baseline controls used in Table 3. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Robustness Check: Human Capital and Education

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.14***	-0.95***	-0.95***	-0.96***	-0.95***	-0.96***
	(0.29)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)
Genetic distance	2.38**	3.13***	3.12***	3.12***	3.10***	3.12***
	(1.05)	(1.18)	(1.18)	(1.18)	(1.18)	(1.18)
Geographic distance	-1.38***	-1.08**	-1.09**	-1.09**	-1.08**	-1.09**
	(0.46)	(0.46)	(0.46)	(0.47)	(0.47)	(0.47)
Human capital index	0.13					
translating country	(0.14)					
Average years of schooling		0.06*				
translating country		(0.03)				
% of primary schooling			-0.01*			0.00
translating country			(0.00)			(0.00)
% of secondary schooling				0.01**		0.01*
translating country				(0.00)		(0.00)
% of tertiary schooling					0.01	0.02
translating country					(0.01)	(0.01)
Benchmark controls	Yes	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,907	8,815	8,815	8,815	8,815	8,815
Adjusted R^2	0.28	0.26	0.26	0.26	0.26	0.26
Country pair clusters	1809	1342	1342	1342	1342	1342

This table tests for selection on the human capital level of a translating country. All regressions include the benchmark set of control variables used in column (6) of Table 3. Country-pair clustered robust standard errors reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A6: Robustness Check: Unobserved Country-Pair Effects

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-0.86*	-0.89*	-0.88*	-0.88*
	(0.47)	(0.52)	(0.52)	(0.52)
Economic controls	No	No	Yes	Yes
Political controls	No	No	Yes	Yes
Trade controls	No	No	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Country Pair FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	42,415	38,928	38,928	38,928
Adjusted R^2	0.36	0.35	0.35	0.35
Country pair clusters	1711	1551	1551	1551

This table tests for selection on time-invariant country-pair effects. Genetic and geographic distance are time invariant across country-pair observations and therefore not estimable in this specification. The set of economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries and the trade controls include the logged value of bilateral trade flows of the country pair. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Robustness Check for Understated Standard Errors

	Dependent variable: Time averaged log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-0.63***	-0.66***	-0.66***	-0.58***	-0.59***	-0.57***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
Genetic distance	1.73***	1.79***	1.78***	1.83***	1.83***	1.69***
	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)
Geographic distance	-0.19***	-0.21***	-0.21***	-0.11	-0.12*	-0.35**
	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.17)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,084	7,084	7,084	7,084	7,084	7,084
Adjusted R^2	0.10	0.11	0.11	0.11	0.11	0.11
Country pair clusters	1891	1891	1891	1891	1891	1891

The dependent variable is the residual of regressing log translations per capita on time dummies and averaged across time. The set of economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries, trade controls include the logged value of bilateral trade flows of the country pair, the colonial controls include a dummy variable indicating if a country pair has ever been in a colonial relationship, and the geography controls include a set of indicators for contiguity and a country pair's absolute difference in latitude and longitude. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Conditional Benchmark Regressions with Cladistic Linguistic Distance

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance (cladistic)	-1.77***	-1.76***	-1.76***	-1.41***	-1.41***	-1.31***
	(0.29)	(0.29)	(0.29)	(0.28)	(0.28)	(0.29)
Genetic distance	2.89***	2.87***	2.87***	3.05***	3.03***	2.16**
	(1.01)	(1.01)	(1.01)	(1.01)	(1.00)	(1.02)
Geographic distance	-0.89***	-0.89***	-0.89***	-0.60***	-0.57***	-1.35***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.17)	(0.45)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39,275	39,275	39,275	39,275	39,275	39,275
Adjusted R^2	0.27	0.27	0.27	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897	1897	1897

This table re-produces the benchmark estimates of Table 3 using a cladistic measure of linguistic distance. The set of economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries, trade controls include the logged value of bilateral trade flows of the country pair, the colonial controls include a dummy variable indicating if a country pair has ever been in a colonial relationship, and the geography controls include a set of indicators for contiguity and a country pair's absolute difference in latitude and longitude. Country-pair clustered robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Robustness Check for Dominant Subject of Translation

	Dependent variable: Log translations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Excluding:	Arts	History	Literature	Philosophy	Religion	Social Science	Applied Science	Natural Science
Linguistic distance	-1.09***	-1.13***	-1.12***	-1.11***	-1.08***	-1.18***	-1.13***	-1.11***
	(0.28)	(0.28)	(0.34)	(0.28)	(0.28)	(0.28)	(0.27)	(0.28)
Genetic distance	2.33**	2.48**	2.47*	2.13**	2.39**	2.78***	2.43**	2.47**
	(1.07)	(1.03)	(1.29)	(1.04)	(1.09)	(1.03)	(1.02)	(1.06)
Geographic distance	-1.34***	-1.32***	-1.54***	-1.42***	-1.36***	-1.44***	-1.33***	-1.33***
	(0.45)	(0.46)	(0.53)	(0.45)	(0.46)	(0.45)	(0.45)	(0.45)
Benchmark controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,405	37,661	26,488	38,610	35,471	37,717	38,374	38,987
Adjusted R^2	0.29	0.30	0.31	0.29	0.30	0.30	0.29	0.28
Country pair clusters	1887	1873	1505	1885	1805	1852	1884	1893

This table reports estimates on various subsamples that exclude each subject classification in the data one by one. All regressions include the benchmark set of control variables used in column (6) of Table 3. Country-pair clustered robust standard errors reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A10: At What Point Does Genetic Distance Become Positive?

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-2.53***	-2.06***	-3.79**	-0.71*
	(0.33)	(0.34)	(1.49)	(0.37)
Genetic distance	-49.31***	-45.01***	3.97***	-2.87
	(7.98)	(8.04)	(1.13)	(4.27)
Geographic distance	-0.84***	-1.40***		
	(0.16)	(0.46)		
Linguistic distance	57.24***	51.77***		
× Genetic distance	(8.92)	(9.00)		
All controls	No	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	39275	39275	26187	13031
Adjusted R^2	0.28	0.29	0.35	0.28
Country pair clusters	1897	1897	1649	681

The interaction estimate in column (2) implies that the marginal effect of genetic distance becomes positive at a threshold of 87 percent dissimilarity in language. In column (3), I report estimates for a subsample of observations above the 87 percent threshold, whereas column (4) reports estimates for a subsample of observations below this threshold. Country-pair clustered robust standard errors in parentheses. All regressions include the full set of control variables used in Table 3. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A11: On the Benefits of Dissimilarity

	Dependent variable: Log translations									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Sample Restrictions:	Full Sample (None)	Linguistic Distance > 10^{th} pctl	Linguistic Distance > 20^{th} pctl	Linguistic Distance > 30^{th} pctl	Linguistic Distance > 40^{th} pctl	Linguistic Distance > 50^{th} pctl	Linguistic Distance > 60^{th} pctl	Linguistic Distance > 70^{th} pctl	Linguistic Distance > 80^{th} pctl	Linguistic Distance > 90^{th} pctl
Genetic distance	-2.99***	-1.07	-0.51	-0.07	0.15	0.51	1.92	3.87**	7.53***	6.08***
	(1.11)	(1.11)	(1.15)	(1.32)	(1.35)	(1.47)	(1.45)	(1.52)	(2.00)	(1.96)
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39275	35426	31276	26825	23428	19227	15554	11694	7878	3982
Adjusted R^2	0.26	0.27	0.3	0.33	0.34	0.36	0.39	0.41	0.43	0.46
Country pair clusters	1897	1832	1767	1662	1551	1390	1219	1005	731	454

This table reports the unconditional estimates of genetic distance used in Figure 3. Column (1) reports full sample estimates, which is identical to column (2) in Table 2. For the estimates in columns (2)-(9), the sample is cut by deciles of linguistic distance, and moving from left to right, the estimates of genetic distance come from increasingly smaller subsamples that are made up of more linguistically distant country-language pairs. Country-pair clustered robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Conditional Baseline Estimates

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.55***	-1.55***	-1.54***	-1.27***	-1.24***	-1.13***
	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)
Genetic distance	3.21***	3.19***	3.19***	3.32***	3.28***	2.40**
	(1.04)	(1.04)	(1.05)	(1.03)	(1.02)	(1.05)
Geographic distance	-0.85***	-0.85***	-0.85***	-0.57***	-0.55***	-1.37***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.17)	(0.46)
Log real GDP per capita (translating)		-0.08	-0.07	-0.17**	-0.18**	-0.15**
		(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Log real GDP per capita (origin)		0.11**	0.11**	0.03	0.02	0.04
		(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Log population (translating)		0.12	0.24	0.38**	0.39**	0.39**
		(0.18)	(0.17)	(0.18)	(0.18)	(0.18)
Log population (origin)		0.09	0.12	0.26**	0.28**	0.30**
		(0.12)	(0.12)	(0.13)	(0.13)	(0.13)
Political rights (translating)			-0.21**	-0.21***	-0.21***	-0.22***
			(0.09)	(0.08)	(0.08)	(0.08)
Political rights (origin)			-0.07	-0.04	-0.03	-0.04
			(0.08)	(0.07)	(0.07)	(0.07)
Log bilateral trade				0.12***	0.13***	0.10***
				(0.02)	(0.02)	(0.02)
= 1 if ever in colonial relationship					-0.13	-0.11
					(0.14)	(0.13)
= 1 for contiguity						0.14*
						(0.08)
Absolute difference in latitude						-0.00
						(0.00)
Absolute difference in longitude						0.01**
						(0.00)
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39275	39275	39275	39275	39275	39275
Adjusted R^2	0.28	0.28	0.28	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897	1897	1897

This table establishes the baseline result for linguistic and genetic distance. Country-pair clustered robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Variable Definitions, Summary Statistics and Data Sources

B.1 Lexicostatistical Linguistic Distance

The computerized approach to estimating lexicostatistical distances was developed as part of the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology. To begin a list of 40 implied meanings (i.e., words) are compiled for each language to compare the lexical similarity of any language pair. Swadesh (1952) first introduced the notion of a basic list of words believed to be universal across nearly all world languages. When a word is universal across world languages, its implied meaning, and therefore any estimate of linguistic distance, is independent of culture and geography. From here on I refer to this 40-word list as a Swadesh list, as it is commonly called.[20]

For each language the 40 words are transcribed into a standardized orthography called ASJPcode, a phonetic ASCII alphabet consisting of 34 consonants and 7 vowels. A standardized alphabet restricts variation across languages to phonological differences only. Meanings are then transcribed according to pronunciation before language distances are estimated.

I use a variant of the Levenshtein distance algorithm, which in its simplest form calculates the minimum number of edits necessary to translate the spelling of a word from one language to another. In particular, I use the normalized and divided Levenshtein distance estimator proposed by Bakker et al. (2009).[21] Denote $LD(\alpha_i, \beta_i)$ as the raw Levenshtein distance for word i of languages α and β. Each word i comes from the aforementioned Swadesh list. Define the length of this list be M, so $1 \leq i \leq M$.[22] The algorithm is run to calculate $LD(\alpha_i, \beta_i)$ for each word in the M-word Swadesh list across each language pair. To correct for the fact that longer words will often demand more edits, the distance is normalized according to word length:

$$LDN(\alpha_i, \beta_i) = \frac{LD(\alpha_i, \beta_i)}{L(\alpha_i, \beta_i)} \tag{2}$$

where $L(\alpha_i, \beta_i)$ is the length of the longer of the two spellings α_i and β_i of word i. $LDN(\alpha_i, \beta_i)$ is the normalized Levenshtein distance, which represents a percentage estimate of dissimilar-

[20] A recent paper by Holman et al. (2009) shows that the 40-item list employed here, deduced from rigorous testing for word stability across all languages, yields results at least as good as those of the commonly used 100-item list proposed by Swadesh (1955).

[21] I use Taraka Rama's (2013) Python program for string distance calculations.

[22] Wichmann et al. (2010) point out that in some instances not every word on the 40-word list exists for a language, but in all cases a minimum of 70 percent of the 40-word list exist.

ity between languages α and β for word i. For each language pair, $LDN(\alpha_i, \beta_i)$ is calculated for each word of the M-word Swadesh list. Then the average lexical distance for each language pair is calculated by averaging across all M words for those two languages. The average distance between two languages is then

$$LDN(\alpha, \beta) = \frac{1}{M} \sum_{i=1}^{M} LDN(\alpha_i, \beta_i). \tag{3}$$

A second normalization procedure is then adopted to account for phonological similarity that is the result of coincidence. This adjustment is done to correct for accidental similarity in sound structure of two languages that is unrelated to their historical relationship. The motivation for this step is that no prior assumptions need to be made about historical versus chance relationship. To implement this normalization the defined distance $LDN(\alpha, \beta)$ is divided by the global distance between two language. To see this, first denote the global distance between languages α and β as

$$GD(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j}^{M} LD(\alpha_i, \beta_j), \tag{4}$$

where $GD(\alpha, \beta)$ is the global (average) distance between two languages excluding all word comparisons of the same meaning. This estimates the similarity of languages α and β only in terms of the ordering and frequency of characters, and is independent of meaning. The second normalization procedure is then implemented by weighting equation (3) with equation (4) as follows:

$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{GD(\alpha, \beta)}. \tag{5}$$

$LDND(\alpha, \beta)$ is the final measure of linguistic distance, referred to as the normalized and divided Levenshtein distance (LDND). This measure yields a percentage estimate of the language dissimilarity between α and β. In instances where two languages have many accidental similarities in terms of ordering and frequency of characters, the second normalization procedure can yield percentage estimates larger than 100 percent by construction, so I divide $LDND(\alpha, \beta)$ by its maximum value to normalize the measure as a continuous $[0, 1]$ variable.

B.2 Cladistic Distance

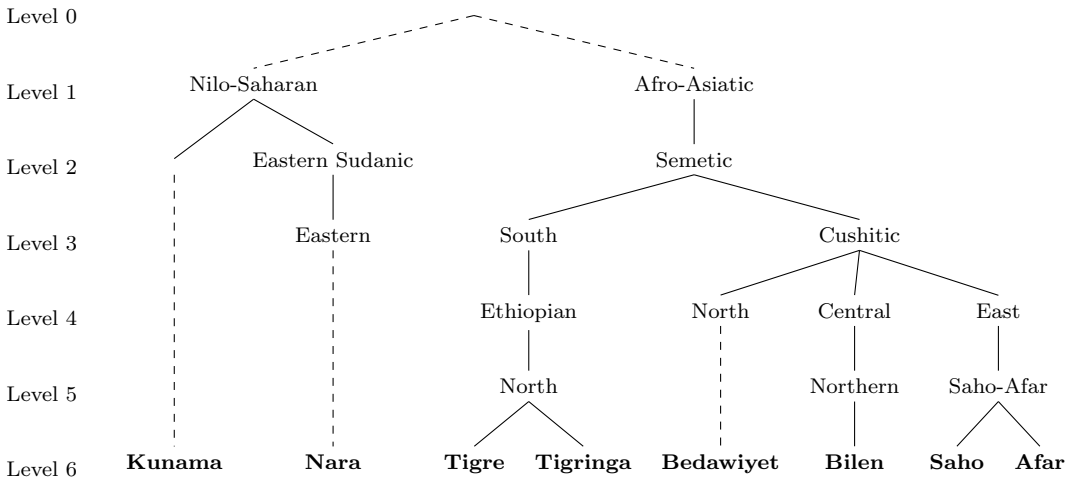
To construct a measure cladistic distance I first calculate the number of shared branches between language α and β on the Ethnologue language tree, denoted $s(\alpha, \beta)$. Let M be the maximum number of tree branches between any two languages. I construct cladistic linguistic distance as follows:

$$CLD(\alpha, \beta) = \left(\frac{M - s(\alpha, \beta)}{M} \right)^{\delta}, \tag{6}$$

where δ is an arbitrarily assigned weight used to discount more recent linguistic cleavages relative to deep cleavages. I describe this weight as arbitrary because there is no consensus on the appropriate weight to be assumed. I follow Fearon (2003) and assume the true function is probably concave with a value of $\delta = 0.5$, although the estimates are robust to alternative weighting scheme used by Desmet et al. (2012).

One issue with calculating cladistic similarity is the asymmetrical nature of historical language splitting. Because the number of branches varies among language families and subfamilies, the maximum number of branches between any two languages is not constant. To overcome this challenge I assume that all current languages are of equal distance from the proto-language at the root of the Ethnologue language tree. I visualize this assumption in Figure B1, where I've constructed a phylogenetic language tree for the 8 distinct languages of Eritrea. The dashed lines represent this assumed historical relationship, so in all cases the contemporary Eritrean languages possess an equal number of branches to the proto-language at Level 0. Although $M = 6$ in Figure B1, in the Ethnologue language tree the highest number of classifications for any language is $M = 15$, which I abstract from here for simplicity.

Figure B1: Phylogenetic Tree of Eritrean Languages



This figure depicts the language tree for the 8 major languages of Eritrea. Because of the asymmetrical nature of language splitting, the number of branches varies among language families. To measure cladistic similarity it is necessary that all branches be extended to the lowest level of aggregation. To do this I assume all languages are of equal distance from the proto-language. Hence, the dashed lines represent the assumed relationship between the proto-language (Level 0) and the set of current Eritrean languages (Level 6).

B.3 Genetic Distance

The F_{st} measure of genetic distance I use is based on an index of heterozygosity – the probability that two randomly selected alleles at a given locus will be different in two populations. An allele distribution that is identical across two populations yields an F_{st} measure equal to zero, while the F_{st} index takes on an increasingly higher value the greater the variation in the allele frequencies across two populations.

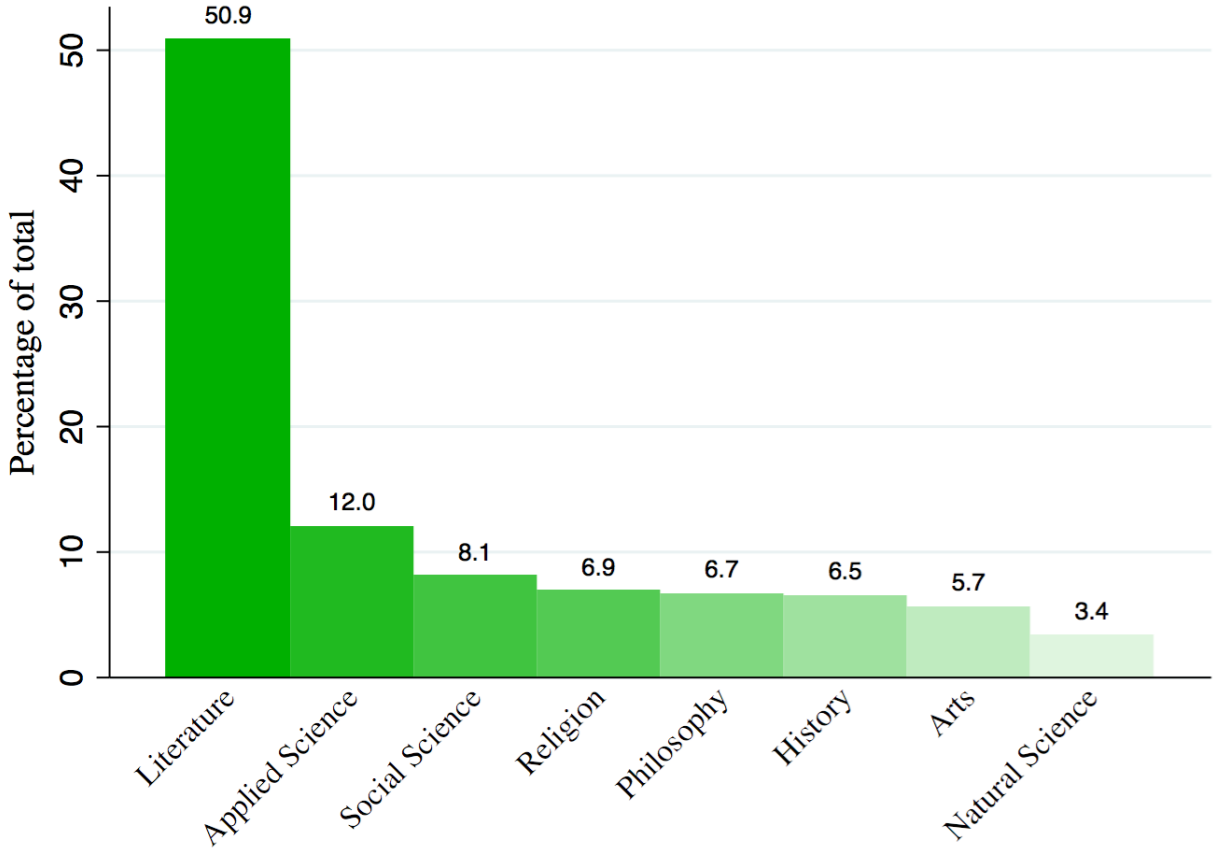
When constructing a measure of genetic distance at the country level it is problematic that many countries contain multiple ethnic sub-groups. To correct for this I adopt a genetic distance measure weighted by the population share of each sub-group in a country.[23] This measures the expected genetic distance between a randomly selected individual from each country. All reported empirical results use this weighted measure versus one that calculates the distance between the dominant population in each country. However, the two measures are highly correlated and the core empirical result is robust to this alternative measure.[24]

[23]For example, suppose country 1 has $i = 1, \ldots, I$ ethnic sub-groups and country 2 has $j = 1, \ldots, J$ ethnic sub-groups with corresponding population shares s_{1i} and s_{2j}. Letting d_{ij} be genetic distance between group i and j, then the weighted F_{st} genetic distance between country 1 and 2 is $F_{st}^{w} = \sum_{i=1}^{I} \sum_{j=1}^{J} (d_{ij} \times s_{1i} \times s_{2j})$.

[24]Because the F_{st} genetic distance data has become more common in the economics literature, I have foregone some details in how the measure is constructed for the sake of brevity. I direct interested readers

B.4 Summary Statistics

Figure B2: Book Translations by Subject



to Spolaore and Wacziarg (2009, p. 480) for a thoughtful and detailed discussion of the data.

Table B1: Benchmark Sample Summary Statistics

	Mean	Std. Dev.	Min.	Max.	N
Log book translations	1.26	1.61	0.00	8.98	39,275
Log economic book translations	0.42	1.01	0.00	7.70	39,275
Log cultural book translations	1.59	1.36	0.00	8.85	39,275
Linguistic distance	0.86	0.11	0.27	1.00	39,275
Linguistic distance (cladistic)	0.96	0.08	0.26	1.00	39,275
Genetic distance	0.04	0.04	0.00	0.29	39,275
Geographic distance (10,000 km)	0.39	0.38	0.01	1.96	39,275
Log real GDP translating country	12.75	1.68	6.17	16.37	39,275
Log real GDP original country	12.97	1.41	5.56	16.26	39,275
Log real GDP original country (weighted)	13.22	1.56	7.12	16.22	37,976
Log population translating country	3.19	1.62	-1.86	7.16	39,275
Log population original country	3.47	1.51	-1.88	7.16	39,275
Log population original country (weighted)	4.13	1.84	-1.86	8.08	37,976
Political rights translating country	0.13	0.24	0.00	0.86	39,275
Political rights original country	0.17	0.28	0.00	0.86	39,275
Political rights original country (weighted)	0.17	0.26	0.00	0.86	37,976
Log bilateral trade	5.88	2.37	-5.91	11.66	39,275
Log bilateral trade (weighted)	5.97	2.31	-5.87	11.66	37,549
= 1 if ever in colonial relationship	0.10	0.30	0.00	1.00	39,275
= 1 if ever in colonial relationship (weighted)	0.08	0.23	0.00	1.00	35,737
= 1 if contiguous	0.13	0.33	0.00	1.00	39,275
= 1 if contiguous (weighted)	0.11	0.30	0.00	1.00	35,737
Absolute difference in latitude	14.94	17.76	0.00	104.2	39,275
Absolute difference in latitude (weighted)	16.14	18.39	0.00	104.2	37,976
Absolute difference in longitude	41.23	44.06	0.07	238.92	39,275
Absolute difference in longitude (weighted)	46.25	46.44	0.02	248.73	37,976
Human capital index translating country	2.60	0.45	1.08	3.54	38,908
Average years of schooling translating country	8.40	2.23	0.62	12.75	8,903
% of primary schooling translating country	22.58	11.76	0.69	55.63	8,903
% of secondary schooling translating country	21.39	13.92	0.48	69.75	8,903
% of tertiary schooling translating country	8.10	4.73	0.14	26.36	8,903

Table B2: **Commonly Translated Authors by Country**

China

Author	Subject
Leo Tolstoy	Literature
Dale Carnegie	Philosophy
Maxim Gorky	Literature

USA

Author	Subject
Rudolf Steiner	Religion
Plato	Philosophy
Anton Chekhov	Literature

Brazil

Author	Subject
Allan Kardec	Philosophy
Joseph Murphy	Religion
Agatha Christie	Literature

Cuba

Author	Subject
Fidel Castro	Social Science
Jose Marti	Literature
Jose Saramago	Literature

Bangladesh

Author	Subject
Syed Abul A'ala Maududi	Religion
Krishna Chandar	Literature
Muhammad Shafi Deobandi	Religion

Saudi Arabia

Author	Subject
Ved Parkash	Literature
Phil Hailstone	Science
Abd Al-Aziz Al-Fawzan	Religion

Argentina

Author	Subject
José Trigueirinho Netto	Philosophy
Sigmund Freud	Psychology
Ramacharaka	Religion

Ethiopia

Author	Subject
Vladimir Lenin	Social Science
William Shakespeare	Literature
Karl Marx	Social Science

Romania

Author	Subject
Nicolae Ceauşescu	Social Science
Ellen Gould White	Religion
Karl Marx	Social Science

Italy

Author	Subject
William Shakespeare	Literature
Fyodor Dostoyevsky	Literature
Augustine of Hippo	Religion

Table B3: Language Pair Observations by Translating Country for the Benchmark Sample (1979-2005)

Country	N	Country	N	Country	N	Country	N	Country	N
Germany	2,330	Brazil	515	Kuwait	159	Mauritius	37	Malawi	8
Spain	2,188	Estonia	442	Argentina	143	Dem. Rep. of Congo	36	Namibia	8
France	2,054	Slovak Republic	426	Indonesia	143	Benin	35	Angola	7
United States	2,007	Greece	424	Sri Lanka	130	Madagascar	34	Botswana	7
India	1,569	Portugal	409	Armenia	127	Luxembourg	32	Panama	6
Switzerland	1,569	Serbia	405	Pakistan	116	Burkina Faso	29	South Africa	6
Sweden	1,275	Lithuania	395	Iran	115	Uruguay	29	Cent. African Rep.	5
Denmark	1,205	Turkey	373	Mongolia	114	Ethiopia	28	Senegal	5
United Kingdom	1,193	Israel	347	Tunisia	105	Malta	28	Swaziland	5
Canada	1,140	Belarus	335	Ukraine	100	Nigeria	27	Cape Verde	4
Belgium	1,073	Croatia	331	Morocco	87	Azerbaijan	26	Ecuador	4
Netherlands	1,008	Slovenia	325	Peru	87	Zimbabwe	26	Mali	4
Finland	986	Iceland	304	Philippines	77	Oman	24	Congo	3
Norway	931	Macedonia	294	Kazakhstan	73	Lebanon	23	El Salvador	3
Japan	884	Albania	292	Cyprus	72	Costa Rica	22	Niger	3
Hungary	870	New Zealand	290	Thailand	72	Cote d'Ivoire	21	Saint Lucia	3
Italy	861	South Korea	282	Iraq	68	Venezuela	15	Kenya	2
Poland	828	Moldova	258	Ireland	64	Nepal	14	Mauritania	2
Russia	802	Mexico	251	Singapore	59	Guatemala	13	Trinidad and Tobago	2
Austria	784	Egypt	229	Bangladesh	55	Qatar	12		
Romania	712	Cameroon	185	Jordon	55	Suriname	12		
Bulgaria	666	Colombia	185	Chad	53	Kyrgyz Republic	11		
Australia	593	Syria	179	Ghana	51	Togo	11		
China	539	Latvia	173	Malaysia	51	Mozambique	10		
Czech Republic	516	Chile	172	Saudi Arabia	45	Bolivia	8		

This table describes the spatial distribution of the benchmark sample (1979-2005). The unit of observation is country-year-language-pair for 119 translating countries, totalling 39,275 observations in the benchmark sample.

Table B4: Observations by Translating Language for the Benchmark Sample (1979-2005)

Translating Language	*N*	Translating Language	*N*	Translating Language	*N*	Translating Language	*N*	Translating Language	*N*	Translating Language	*N*	Translating Language	*N*
English	5,538	Latvian	213	Tagalog	33	Waama	11	Mari	7	Hdi	5	Biali	3
German	3,099	Urdu	208	Irish	30	Aragonese	10	Morisyen	7	Jula	5	Fijian	3
French	3,054	Latin	175	Northern Saami	29	Bissa	10	Pitjantjatjara	7	Kabiye	5	Igbo	3
Spanish	2,297	Mongolian (Halh)	175	Maltese	28	Kera	10	Somrai	7	Karachay-Balkar	5	Kabardian	3
Dutch	1,370	Malayalam	169	Maori	28	Koonzime	10	Tuva	7	Karelian	5	Moksha	3
Italian	1,242	Persian (Iranian)	168	Western Frisian	26	Lamnso	10	Arabic (Chadian)	6	Kekchi	5	St. Lucian Creole	3
Russian	1,057	Galacian	167	Kalaallisut	26	Yoruba	10	Blaan, Sarangani	6	Kenga	5	Sranan	3
Portuguese	1,014	Idonesian	163	Occitan	25	Hausa	9	Bosnian	6	Kituba	5	Suri	3
Swedish	1,005	Tamil	151	Tatar	25	Kom	9	Chuvash	6	Koorete	5	Udmurt	3
Arabic	968	Armenian	150	Georgian	24	Kuo	9	Corsican	6	Koromfe	5	Veps	3
Japanese	934	Ukrainian	142	Kasem	24	Ladino	9	Fon	6	Manobo, Cotabato	5	Aramaic	2
Hungarian	926	Belarusan	126	Uzbek	21	Northern Ndebele	9	Haitian	6	Naro	5	Buriat	2
Norwegian	891	Sinhala	117	Samoan	20	Niuean	9	Kabyle	6	Ngangam	5	Chukchi	2
Polish	870	Vietnamese	111	Sanskrit	20	Ossetian	9	Kalagan	6	Carpathian Romani	5	Coptic	2
Romanian	794	Esperanto	109	Central Tibetan	20	Plautdietsch	9	Limbum	6	Syriac	5	Cornish	2
Danish	793	Kazakh	109	Kako	19	Chichewa	8	Luba-Lulua	6	Turkmen	5	Dargwa	2
Finnish	753	Assamese	106	Tongan	18	Chipewyan	8	Mampruli	6	Vengo	5	Even	2
Bulgarian	683	Oriya	93	Amharic	16	Daba	8	Mapun	6	Warlpiri	5	Frisian	2
Czech	602	Breton	85	Luxembourgeois	16	Dangaleat	8	Nafaanra	6	Aghem	4	Guadeloupean Creole	2
Catalan	488	Marathi	84	Tigrigna	16	Gaelic	8	Nepali	6	Aja	4	Guarani	2
Greek	488	Kannada	77	Tokelauan	16	Hawaiian	8	Noone	6	Bamanankan	4	Jingpho	2
Mandarin	455	Gujarati	76	Cree	15	Khmer	8	Ojibwa	6	Old Church Slavic	4	Kara-Kalpak	2
Turkish	437	Faroese	75	Kyrgyz	15	Konkomba	8	Roviana	6	Djeebbana	4	Maithili	2
Albanian	369	Thai	75	Mofu-Gudur	15	Lao	8	Sisaala, Tumulung	6	Hmong	4	Navajo	2
Estonian	359	Telugu	73	Rarotongan	15	Lingala	8	Swati	6	Komi-Zyrian	4	Ndonga	2
Hebrew	356	Uyghur	69	Scots	15	Mundani	8	Tboli	6	Kongo	4	Nogai	2
Lithuanian	356	Inuktitut	65	Swahili	15	Nateni	8	Yakut	6	Kriol	4	Reunion Creole	2
Croatian	352	Welsh	60	Tajiki	15	Newari	8	Zulgo-Gemzek	6	Mukulu	4	Southern Saami	2
Korean	330	Malay	56	Swabian	14	Parkwa	8	Akoose	5	Balkan Romani	4	Southern Sotho	2
Slovene	321	Serbo-Croatian	49	Kankanaey	13	Yamba	8	Anyin	5	Macedo Romanian	4	Tahitian	2
Slovak	314	German (Swiss)	45	Farefare	12	Afrikaans	7	Bafut	5	Inari Saami	4	Tai Hongjin	2
Serbian	309	Kurdish	44	Tikar	12	Bashkir	7	Chumburung	5	Lule Saami	4	Tamazight	2
Icelandic	296	Panjabi	43	Gude	11	Fuliiru	7	Dakota	5	Sokoro	4	Tok Pisin	2
Macedonian	240	Asturian	42	Karang	11	Gbaya-Bossangoa	7	Dan	5	Tswana	4		
Bengali	234	Azerbaijani	37	Ngbaka	11	Kalinga	7	Gagauz	5	Wolof	4		
Hindi	230	Yiddish	37	Shona	11	Kenyang	7	Gikyode	5	Adyghe	3		
Basque	213	Malagasy	33	Somali	11	Mambila	7	Hanga	5	Avar	3		

This table reports the benchmark sample by translating language (1979-2005). The unit of observation is country-year-language-pair for 119 translating countries, totalling 39,275 observations in the benchmark sample.

Table B5: Observations by Original Language for the Benchmark Sample (1979-2005)

Original Language	N	Original Language	N	Original Language	N	Original Language	N	Original Language	N	Original Language	N
English	4,958	Korean	282	Mongolian (Halh)	59	Ladino	17	Friulian	6	Shona	4
French	3,066	Albanian	264	Malay	54	Lao	17	Guarani	6	Abkhazian	3
German	2,770	Hindi	254	Telugu	54	Uzbek	17	Lingala	6	Asturian	3
Greek	2,349	Croatian	252	Galacian	52	Zulu	17	Yucatan Maya	6	Chechen	3
Spanish	1,544	Vietnamese	204	Quiche	51	Bamanankan	16	Huautla Mazatec	6	Chukchi	3
Dutch	1,255	Slovene	195	Marathi	50	Geez	16	Southern Sotho	6	Erzya	3
Arabic	1,241	Pali	186	Belarusan	49	Maltese	16	Udmurt	6	Northern Frisian	3
Swedish	1,214	Slovak	176	Inuktitut	45	Avestan	14	Cree	5	Ganda	3
Russian	1,115	Estonian	163	Malayalam	43	Dakota	14	Hopi	5	Hmong	3
Danish	1,066	Indonesian	160	Amharic	42	Oriya	14	Igbo	5	Kabardian-Cherkess	3
Portuguese	1,063	Syriac	155	Swahili	41	Tagalog	14	Kalaallisut	5	Kongo	3
Japanese	1,053	Panjabi	139	Nepali	40	Javanese	13	Komi-Zyrian	5	Maori	3
Hebrew	1,009	Irish	138	Faroese	39	Turkmen	13	Manchu	5	St. Lucian Creole	3
Polish	991	Arabic (Egyptian)	136	Breton	35	Western Frisian	11	Moksha	5	Zarma	3
Mandarin	933	Ukrainian	130	Gaelic	34	Gikuyu	11	Morisyen	5	Arabic (Moroccan)	2
Norwegian	902	Aramaic	127	Gujarati	33	Tajiki	11	Plautdietsch	5	Avar	2
Hungarian	776	Afrikaans	126	Azerbaijani	29	Wolof	11	Lule Saami	5	Chagatai	2
Persian (Iranian)	655	Lithuanian	115	Old Church Slavic	24	Carib	10	Sranan	5	Fang	2
Finnish	640	Coptic	110	Northern Saami	24	Uyghur	9	Cornish	4	Kalmyk	2
Romanian	599	Thai	109	Kazakh	23	Cheyenne	8	Corsican	4	Khanty	2
Turkish	571	Latvian	104	Tamazight	22	Evenki	8	Duala	4	Koryak	2
Sanskrit	519	Armenian	92	Kyrgyz	21	Hausa	8	Komi-Permyak	4	Lak	2
Bengali	493	Tamil	88	Sorbian	21	Sinte Romani	8	Luxembourgeois	4	Mansi	2
Czech	490	Kurdish	85	German (Swiss)	20	Tatar	8	Mari	4	Mapudungun	2
Bulgarian	399	Occitan	83	Scots	20	Assamese	7	Moore	4	North Ndebele	2
Tibetan	388	Welsh	82	Tamasheq	19	Chuvash	7	Navajo	4	Ossetian	2
Yiddish	378	Basque	78	Kannada	18	Kashmiri	7	Nenets	4	Rwanda	2
Icelandic	368	Macedonian	74	Malagasy	18	Vlax Romani	7	Ojibwa	4	Seraiki	2
Urdu	334	Esperanto	68	Sinhala	18	Akan	6	Carpathian Romani	4	Tuva	2
Catalan	327	Georgian	59	Yoruba	18	Ewe	6	Macedo Romanian	4		

This table reports the benchmark sample by translating language (1979-2005). The unit of observation is country-year-language-pair for 119 translating countries, totalling 39,275 observations in the benchmark sample.

B.5 Data and Sources

Book translations: The total number of translated books in a country for a given year. The data used here is from the time period 1979-2005, and comes from the Index Translationum, an online bibliographic archive hosted by UNESCO.
Source: http://www.unesco.org/culture/xtrans/

Lexicostatistical linguistic distance: This computerized lexicostatistical linguistic distance measures the phonetic similarity between two languages. See Appendix B for a formal discussion of how this data is estimated. I source the language lists used to estimate linguistic distance from the Automated Similarity Judgement Program (ASJP) to estimate language distances (Wichmann et al., 2013).
Source: http://asjp.clld.org/

Cladistic linguistic distance: I construct the cladistic measure according to the details above in Appendix B. I use the Ethnologue language tree, sourced from World Language Mapping System.
Source: http://www.worldgeodatasets.com/language/

Genetic distance: An index of heterozygosity, the probability that two randomly selected alleles at a given locus will be different in two populations. Genetic distance is used as a proxy for the degree of common ancestry between two populations. This data was originally constructed by Cavalli-Sforza et al. (1994), and was sourced from Spolaore and Wacziarg (2009).
Source: http://www.anderson.ucla.edu/faculty_pages/romain.wacziarg/downloads/genetic_distance.zip

Geographic distance: Geodesic distance between the most populated cities in a country pair, measured per 100 kilometres. This data comes from a data set compiled by researchers at Centre d'Etudes Prospectives et d'Informations Internationales (CEPII).
Source: http://www.cepii.fr/cepii/en/welcome.asp

Real GDP: Purchasing power parity converted expenditure-side real GDP at chained purchasing power parity rates (in mil. 2005 US dollars) from the Penn World Tables version 8.0.
Source: http://www.rug.nl/research/ggdc/data/pwt/pwt-8.0

Population: Total population in thousands from the Penn World Tables version 8.0.
Source: http://www.rug.nl/research/ggdc/data/pwt/pwt-8.0

Political rights: Freedom House Political Rights Index with an original range of 1 through 7, normalized as a 0-1 variable.
Source: http://www.freedomhouse.org/report-types/freedom-world#.U1f1lV5bTwI

Colonial history indicator: Colonial history data was sources from the CEPII. I use an dummy variable indicating if a country pair has ever been in a colonial relationship.
Source: http://www.cepii.fr/cepii/en/welcome.asp

Other geography data: All other geography data was also sourced from the CEPII. Measure of latitude and longitude differences were constructed using individual country measures and taking the absolute value of the difference between each for a country pair. A dummy variable was also collected indicating contiguity of a country pair.
Source: http://www.cepii.fr/cepii/en/welcome.asp

Shared common language indicators: Indicator variables specifying (i) if country pairs share a common language and (ii) if at least 9 percent of the population in both countries speak the same language were also sourced from the CEPII.
Source: http://www.cepii.fr/cepii/en/welcome.asp

Bilateral trade shares: Measures the logged average value of bilateral trade for a country pair in constant US dollars. To construct this variable I average imports from i to j and imports from j to i in current US dollars, deflate this value by the American CPI for all urban consumers (1982-1984 = 100; taken from http://www.bls.gov/data/#prices), and take the log of this averaged value. The trade data was sourced from Barbieri et al. (2009).
Source: http://www.correlatesofwar.org/COW2%20Data/Trade/Trade.html

Human capital: The cross-country 5-year panel of education attainment for the total population aged 15 and over is from Barro and Lee (2013). Measures used include (i) the average years of schooling attained, (ii) the percentage of complete primary schooling attained, (iii) the percentage of complete secondary schooling attained, and (iv) the percentage of complete tertiary schooling attained.
Source: http://www.barrolee.com

Human capital index: The index of human capital per person is based on years of schooling (Barro and Lee, 2013) and returns to education (Psacharopoulos, 1994), and is sourced from the Penn World Tables version 8.0.
Source: http://www.rug.nl/research/ggdc/data/pwt/pwt-8.0